GENDER DIFFERENCES ON THE REVISED SOCIOSEXUAL ORIENTATION INVENTORY: A DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

A THESIS

SUBMITTED TO THE GRAUDATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

BY

KIM KIRKEBY

DR. HOLMES FINCH-ADVISOR

BALL STATE UNIVERSITY

MUNCIE, INDIANA

JULY 2019

ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my committee chair, Dr. Finch Holmes, for his guidance. I am extremely grateful for the expertise and patience that he provided throughout the duration of this project—particularly since the chosen topic for this work was, at first, far outside my area of comfortability.

A sincere thank you also goes to Dr. Mary Kite and Dr. Jocelyn Bolin for sharing their expertise and offering such thoughtful feedback. Their insight has undoubtedly greatly improved the quality of this work.

Last but not least, I would like to acknowledge my parents, Mary Sandoval and Kevin Kirkeby. Without their continued support and encouragement, my academic goals would not have become a reality. I am eternally grateful for all they have done to help me succeed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I: INTRODUCTION

Arguably, one of the most enduring debates is the nature of gender differences, why they exist, and even if they really exist at all (Peterson & Hyde, 2010; Wood & Eagly, 2002). Numerous theories have been posed over the years ascribing observed differences between the sexes including evolutionary adaptations (Buss, 1995), social cognitive theories (Bandura, 1986), social roles held in society (Wood & Eagly, 2012), and that the two genders are more similar than they are different (Hyde, 2005). Regardless, one of the most consistent assertions is that the largest disparities between men and women can be observed in the area of sexual behaviors and attitudes (Peterson & Hyde, 2010; Schmitt, 2005). Throughout the literature, men are reported as desiring sex more often, with a larger variety of partners, and having a more favorable attitude toward sex outside the context of a committed relationship (Penke & Asendorpf, 2008; Schmitt, 2005; Simpson & Gangestad, 1991). Although these findings are consistent on self-report assessments examining sexual behavior and attitudes, they rely heavily on the presumption that the items within these instruments demonstrate validity by behaving the same way for both men and women. In some cases, however, such a presumption may prove to be inaccurate given the well-documented tendencies of individuals—especially women—to respond to items pertaining to sexual behavior and attitudes in a socially desirable manner that conforms to gendered norms and social expectations (Alexander & Fisher, 2003; Fenton, Johnson, McManus, & Erens, 2001; Mitchell et al., 2018). Consequently, instruments assessing these constructs may have items that perform differently by gender, specifically by being easier for men to endorse than they are for women. Such differences in item behavior are known as differential item functioning (DIF; Zumbo, 1999). Because the assurance that test items are equally valid for all subgroups of a population is crucial, determining whether items exhibit DIF has important implications for the accuracy of claims made regarding personality traits and gender differences. As such, the goal

of the proposed study is to assess whether DIF is present in the items comprising an instrument commonly used in personality and sexuality research, known as the revised Sociosexual Orientation Inventory (SOI-R, Penke & Asendorpf, 2008), using the multiple indicators, multiple causes (MIMIC) model (Jöreskog & Goldberger, 1975).

CHAPTER 2: LITERATURE REVIEW

Sociosexuality

Sociosexuality (also referred to as sociosexual orientation) is a construct that describes individual differences in the willingness to engage in uncommitted sexual relations, colloquially known as "casual sex" (Kinsey, Pomeroy, & Martin, 1948; Penke & Asendorpf, 2008; Simpson & Gangestad, 1991). Sociosexuality is measured on a continuum, where lower scores indicate a "restricted" orientation characterized by a preference for engaging in sexual intimacy in the context of a committed relationship, whereas higher scores indicate an "unrestricted" orientation characterized by a preference for a variety of short-term sexual partners. Individuals with restricted orientations report having had few sexual partners during the past year, few to no instances of having sex with a partner on only one occasion, and discomfort with sex prior to developing emotional closeness with potential partners. Conversely, individuals with unrestricted orientations report having had a variety of sexual partners during the past year, are able to easily engage in sexual encounters without emotional closeness and tend to have had numerous occasions of engaging in sex with a partner only once (Simpson & Gangestad, 1991).

Evolutionary theory proposes that these preferences have evolved as a result of the various reproductive challenges that ancestral men and women faced throughout history (Simpson & Gangestad, 1991). In short, it is theorized that women have evolved to preference long-term mating strategies with the goal of retaining a mate who can invest resources in offspring whereas men have evolved to maximize the likelihood of passing on their genetics by

mating with a greater number of women short-term (Buss, 1998; Buss & Schmitt, 1993; Trivers, 1972). This is thought to explain between gender differences in sociosexuality, as an unrestricted orientation is associated with greater engagement in short-term mating behaviors—which is more prevalent in men— and a restricted orientation is associated with long-term mating behaviors—which is more prevalent in women (Buss, 1998; Penke & Asendorpf, 2008; Schmitt, 2005; Simpson & Gangestad, 1991). It is important to note, however, that greater individual differences (within gender) in sociosexuality have been observed than those between the two genders, which is theorized to be a result of the reproductive advantages afforded to men and women who had the ability to adapt their mating strategies with environmental and cultural circumstances (Buss, 1998; Simpson & Gangestad, 1991). Although reproduction is no longer contingent upon these dimensions in the modern world, it is thought that humans retain these artifacts of their evolutionary past through their desires, motivations, and traits, such as sociosexuality. Furthermore, these models offer an explanation as to why men express stronger sexual desire for a variety of sexual partners and generally exhibit more permissive attitudes toward uncommitted sex than do women. Having an unrestricted orientation has been linked to traits such as extraversion, openness to experience, sensation seeking, and erotophilia (Simpson, Wilson, & Winterheld, 2004).

Initially, sociosexuality was measured as one unidimensional trait using the 7-item instrument known as the Sociosexual Orientation Inventory (SOI; Simpson and Gangestad, 1991). Although the SOI became widely used, measuring such a complex construct as a sociosexuality as a unidimensional trait based soley on sexual behavior was heavily criticized, as there were thought to be several extraneous factors that might influence behavior (Penke & Asendorpf, 2008). For example, environmental factors such as social norms, religion, being in a

long-term relationship, or even an individual's ability to access potential mates could have a restrictive effect on sexual behaviors and, to a lesser extent, attitudes towards sex. Yet, sexual desire for a variety of short-term mates may remain strong, underscoring the need for an instrument that assessed sociosexuality as a multidimensional construct. Penke and Asendorpf (2008) theorized that sociosexuality was comprised of three latent traits: a component consisting of past sexual behavioral experiences, attitudes toward casual sex, and the desire to engage in casual sex. In response, a revised version of the Sociosexual Orientation Inventory (SOI-R) that reflected these three traits was created and is the instrument of focus for the proposed study.

In general, men typically score higher than women on the SOI-R, most consistently on the desire facet, which is thought to be shaped more by biological factors. In contrast, sexual attitudes and behaviors are thought to be more heavily influenced by social and cultural factors (Penke & Asendorpf, 2008; Schmitt, 2005; Simpson & Gangestad, 1991). Although such gender differences are consistently found and attributed to actual differences between men and women, research suggests that response patterns may be different for men and women and are thus prone to inaccuracies (Fenton et al., 2001; Krumpal, 2013).

Challenges in Measuring Sexual Attitudes and Behaviors

Measuring attitudes towards and experiences with sexual behaviors presents special challenges when it comes to obtaining accurate and unbiased reports from respondents (Alexander & Fisher; 2003; Krumpal, 2013; Mitchell et al., 2018). This is largely because, unlike other commonly studied behaviors, sexual behavior is a private activity that tends to be constrained by social, cultural, religious, legal, and moral norms and therefore is a sensitive topic (Fenton et al., 2001). People, in general, are inclined to employ impression management strategies to appear favorably in front of others (Baumeister & Finkel, 2010; Schlenker, 1980). As a result, respondents may be particularly unwilling to report truthfully on items inquiring

about attitudes or behaviors where they perceive that they may be judged for their response, such as those pertaining to sexuality. Similarly, items inquiring about sensitive issues may elicit social desirability in responding, which refers to the tendency for people to respond or present themselves in a positive light, regardless of their actual behavior or attitudes (Krumpal, 2013). In particular, people tend to underreport attitudes or behaviors that society deems undesirable and overreport those that are socially desirable (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Fenton et al., 2001; Krumpal 2013).

One of the most consistent finding in the sexuality literature is that men tend to overreport numbers of past sexual partners and women tend to underreport theirs (Alexander &Fisher, 2003; Fenton et al., 2001; Mitchell et al., 2018). While it may be impossible to determine the exact number of sexual partners that men or women have had over their lifetimes, they should at least match up mathematically, given that whenever a man has sex with a woman, that woman is also having sex with a man. One explanation for the observed discrepancy in reported numbers of sexual partners is that of gender stereotypes, which state that men are expected to adopt agentic, dominant roles, be sexually active, and the initiators of sexual behavior. Conversely, women are expected to adopt communal roles, be submissive and reactive to men's sexual advances, and exercise restraint over their own sexual desires and behaviors (Emmerink, Vanwesenbeeck, van den Eijnden, & ter Bogt, 2016). Consequently, there is general social acceptance of men when they engage in sex outside of the context of a committed relationship with a variety of women; however, the same behavior is generally not accepted when enacted by a woman in a pattern known as the (hetero) sexual double standard (SDS). When members of either gender violates these norms, they are likely to be subjected to consequences, such as losing social status, or being the victims of rumors and gender-based

harassment. Although some research suggests a loosening of the SDS in recent decades, and that women's self-reported attitudes toward sexuality have grown more liberal, other findings indicate that there has been little change (Marks & Fraley, 2005; Peterson & Hyde, 2010; Sprecher, Treger, & Sakaluk, 2013; Zaikman & Marks, 2017). As a result, many of the items used to assess sociosexuality may not fully reflect actual gender differences, but rather false accommodations to reflect responses consistent with these gendered norms. Indeed, the findings from recent research suggests that gender discrepancies in the reporting of lifetime sexual partners can be explained by a few key differences between women and men (Mitchell et al., 2018). First, men and women utilize different counting strategies when it comes to past sexual partners. Whereas men tend to estimate when accounting past partners, women are more likely to methodically count. Second, women are more likely than men to include partners with whom they have only engaged in oral sex. Third, women hold more conservative views toward casual and nonexclusive sex, both of which predict reporting only one sexual partner during the past year. Finally, a small percentage of men report extreme values of past sexual partners. The researchers adjusted for these discrepancies by capping partner numbers in the 99th percentile and adjusting for both counting strategies and sexual attitude differences. Upon doing so, gendered differences narrowed substantially. Although some research suggests that anonymously administering questionnaires pertaining to sensitive topics helps to curb social desirability in responding, doing so does not fully eliminate these biases (Dodou & de Winter, 2014). Clearly, this presents a problem, as instruments utilized in research are expected to accurately measure the latent construct free from systematic error (Clauser & Mazor, 1998; Podsakoff et al., 2003).

Previous research also indicates that endorsement of the SDS is associated with greater adherence to traditional gender role norms, which in turn predicts false accommodation of

responses on items regarding sexual attitudes and behaviors (Mitchell et al., 2018). Although findings are somewhat mixed, empirical research suggests that men are more likely to endorse the SDS (presumably because they benefit from it more than women), and women may experience greater pressure to conform to gendered norms of sexual behaviors and attitudes (Emmerink et al., 2016; Zaikman & Marks, 2017). Interestingly, having an unrestricted orientation is positively correlated with endorsing traditional gendered norms for men, but negatively correlated for women, which could also influence gender differences in item performance (Simpson et al., 2004). In short, this evidence suggests that items assessing sexual behaviors and attitudes may be performing differently for different groups (i.e., men and women).

Such differences in item performance are inherently problematic because there is an expectation that an instrument is basing respondents' scores soley on the latent trait of interest and that the instrument is performing equally for all subgroups of the population that have identical scoring on measures of the latent trait (Clauser & Mazor, 1998). In other words, if men and women are matched on sociosexuality, they should have equivalent probabilities of endorsing a response on the SOI-R. For example, when men and women are matched on the latent trait, it should not be easier for men to agree with Item 4 of the SOI-R, "*Sex without love is ok*." In the event that an item is easier for a man to endorse than for a woman, that item can be said to behave differently across groups. This is known as differential item functioning, or DIF (Zumbo, 1999).

Although a review of the literature did not reveal any previous research assessing DIF in instruments assessing sociosexuality, DIF has been noted in other personality assessments. For example, De Leo, Van Dam, Hobkirk, and Earleywine (2011) found both uniform and non-

uniform DIF for several of the items of the Impulsive Sensation Seeking scale on several sociodemographic variables, including gender. Similarly, several items of widely used personality inventories, such as the abridged Big-Five Circumplex (Mitchelson, Wicher, LeBreton, Craig, 2009) show gender and ethnicity DIF on an inventory that is often used in hiring considerations. These findings are theorized to be a result of socio-cultural factors shaping what traits women versus men and members of white communities versus black communities place value on, thereby influencing how different groups responded to items.

Likewise, socio-cultural factors may also be influencing responses to the SOI-R, as evidenced by the aforementioned literature indicating that responses to sexuality surveys are prone to false accommodation to appear in line with social norms—which may be particularly prevalent among women. Because this suggests that many (if not all) of the items of the SOI-R may exhibit DIF, the goal of this study was to perform a DIF analysis on all items of the scale. Currently, there are many statistical approaches used to detect DIF, with many being based in item response theory (IRT) methodology (de Ayala, 2009). IRT provides a framework for describing the relationship between an observed response on a given set of items and the respondent's levels of the latent trait measured by the item set (Lord & Novick, 1968). Because an understanding of IRT methodology is useful in the interpretation and understanding of DIF, a summary of IRT and IRT models will be described in the following paragraphs. A description of DIF will also be provided, as well as a description of some common procedures for DIF detection.

Item Response Theory

IRT is an approach to educational and psychological measurement that focuses on two distinct components of measurement: an individual's latent abilities and the characteristics of test items (Lord & Novick, 1968). By assessing the relationship between these two components, the

probability of a correct response to an item can be estimated. The overarching goal with IRT is to assess the probability that an individual will provide a correct response given that individual's latent ability (denoted as θ) and item characteristics (item difficulty, discrimination, and pseudo-guessing). In IRT, θ is a standardized representation that depicts how far from average an individual's ability level on the latent trait is, where zero represents average θ, negative values indicate lower than average θ, and positive values indicate higher than average θ. Additionally, IRT is also concerned with the item characteristics of difficulty and discrimination, both of which are also centered at zero, with negative values indicating easier items and positive values indicating items that are more difficult. With IRT, individual ability is incorporated into the interpretation of item difficulty, such that correct responses to more difficult items require higher ability levels. In general, items with discrimination parameter values higher than one are considered good discriminators, meaning that the items are adept at differentiating between individuals with high (in this case unrestricted orientations) and low latent ability (restricted orientations), whereas a discrimination value less than .8 is considered low.

To calculate θ, there are several methods available; however, the most popular utilizes maximum likelihood estimation (MLE), which calculates the probability of obtaining the response pattern observed in the data in a process known as the likelihood function. MLE is expressed as:

$$L(X_i|\theta, b) = \prod_{j=1}^{J} P_j^{X_{ij}} \left(1 - P_j\right)^{\left(1-X_{ij}\right)} \quad (1)$$

Where $P_j^{X_{ij}}$ is the probability of the given response to item j given item response pattern X_{ij}. For example, an examinee who provides correct responses to four items on an exam, and misses the last item, would have a response pattern of 1110. Essentially, MLE finds the value of θ that would make the observed responses most likely, taking the difficulty of each item into account.

IRT utilizes item characteristic curves (ICC) in interpreting item functioning and depicts the relationship between θ and the probability of endorsing the item (Lord, 1952). The position of the ICCs on the X-axis (the latent variable) and the Y-intercept (probability of a correct response) provide a visual representation of this relationship. Similarly, item information curves (IIC) are used to depict item discrimination. Steeper curves represent items with higher discrimination values, thus providing more information on a respondent's ability whereas flat, broad curves indicate items with poor discrimination and provide less information.

There are two broad categories of IRT models: those that utilize dichotomous items, meaning that the item has two response categories that are scored as correct versus incorrect, and those that utilize polytomous items, which have three or more categories (de Ayala, 2009). One, two, and three parameter logistic models are fit to the data and are based on how many of the parameters are utilized to estimate the relationship between ability and item response patterns. The most commonly utilized IRT models for dichotomous include the Rasch, the 1 parameter logistic model (1PL), the 2PL, and the 3PL—all of which have counterpart models for use with polytomous items and are summarized in the following paragraphs.

Dichotomous IRT Models

Rasch model. The Rasch model is the simplest of IRT models and estimates the probability of a correct response from person ability and item difficulty (de Ayala, 2009). Discrimination is constrained to one for all items so that all items are assumed to discriminate among respondents equally, with one generally considered to be a good discrimination value. The Rasch model is expressed as:

$$P(x_j = 1|\theta, b) = \frac{e^{(\theta_i - b_j)}}{1+e^{(\theta_i - b_j)}} \quad (2)$$

Where θ_i represents the ability level for person *i*, and b_j represents difficulty for item *j*

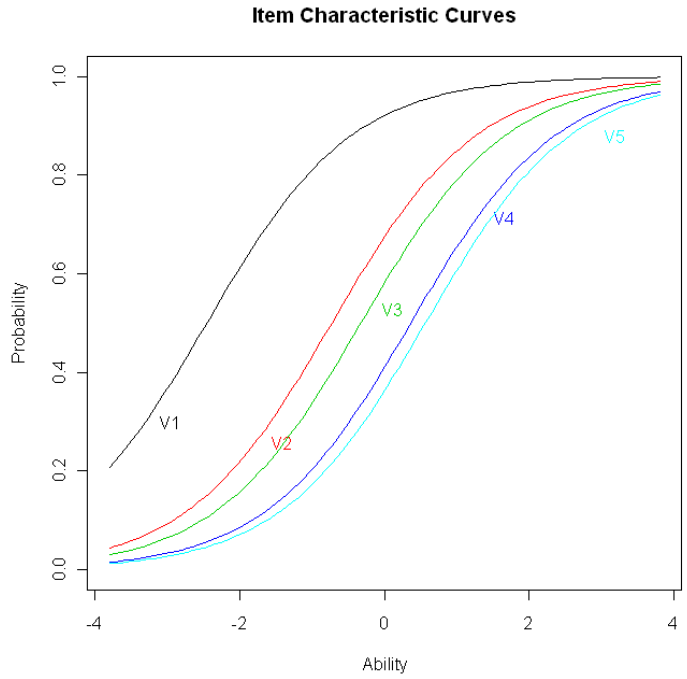


Figure 1.1 Item Characteristic Curves for the Rasch Model.

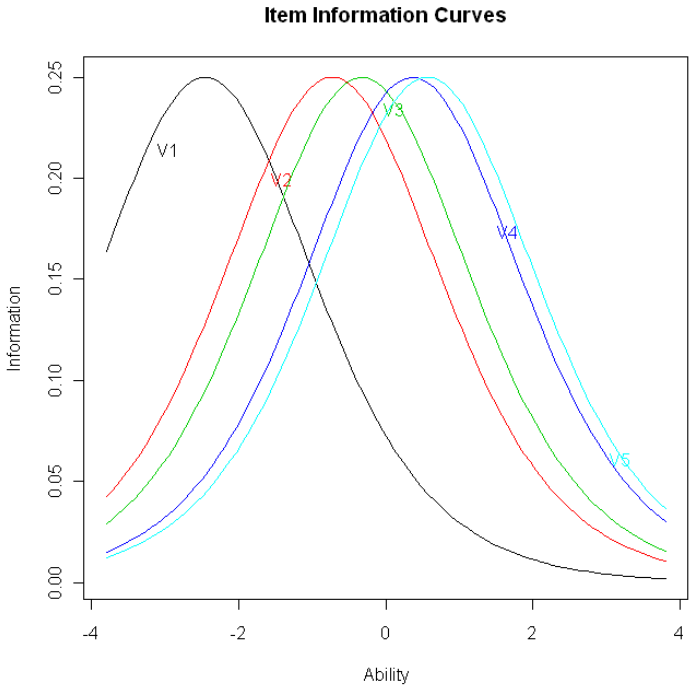


Figure 1.2. Item Information Curves for the Rasch Model

As depicted in Figure 1.1, the slope for all five of the items are the same, indicating that the discrimination parameters are identical. The plot shows that item difficulty is lowest for Item V1, as indicated by its position on the X axis, which is furthest to the left. The ICC for Item V1 shows that individuals on the lower end of the ability continuum have a probability of getting the item correct of approximately 20% (.2) and a probability of about 80% for those with average θ . The location of Item V5 indicates that it is a more difficult item given that an individual must have average θ to have a 20% probability of getting the item correct. The plot for item information curves (IIC) for the Rasch model (Figure 1.2) also illustrates identical discrimination parameters for each item when constrained to a value of one. The position of the peaks on the X-axis represent θ for which the maximum amount of information is provided by each item. As can be seen, Item V1 provides the most information on respondents with lower θ, items V2 and V3 provide the most information for those with average θ, and items V4 and V5 provide the most information for those with θ just above average.

1 PL. The one parameter logistic model (1PL) is similar to the Rasch in that discrimination across all items is estimated as a single value; however, this value is not constrained to one. The 1PL can be expressed as:

$$P(x_j = 1|\theta, b, a) = \frac{e^{a(\theta_i - b_j)}}{1+e^{a(\theta_i - b_j)}} \quad (3)$$

Where θ_i denotes ability level for person *i*; b_j represents difficulty for item *j*; and a represents discrimination for all items. As can be seen in Figure 2.1, the steepness of the ICC is slightly higher than that observed in Figure 1.1 as a result of the discrimination parameter being estimated at a higher at a value of 1.4 when fit with the 1PL model. Similar to the Rasch model, however, the slopes for each item are identical. Although steeper, the IIC plot for the 1PL model (Figure 2.2) is also nearly identical to that seen with the Rasch model.

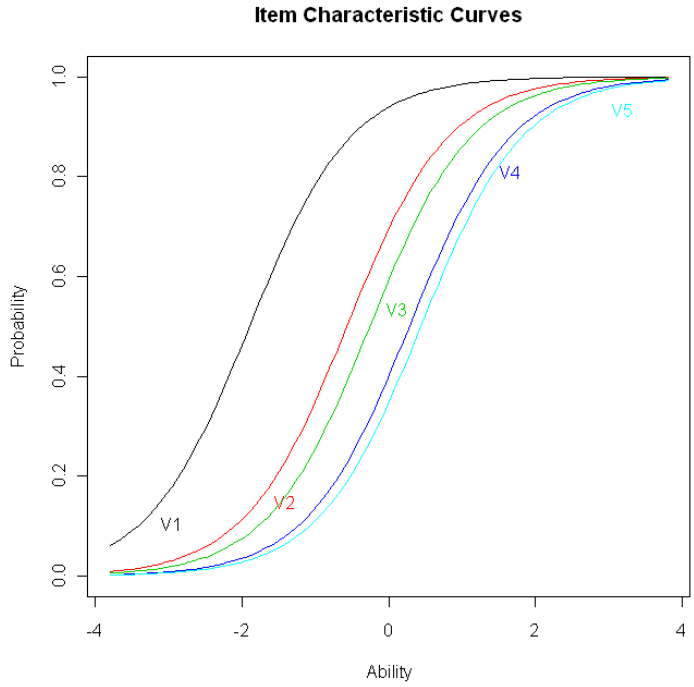


Figure 2.1. Item Characteristic Curves for the 1PL Model

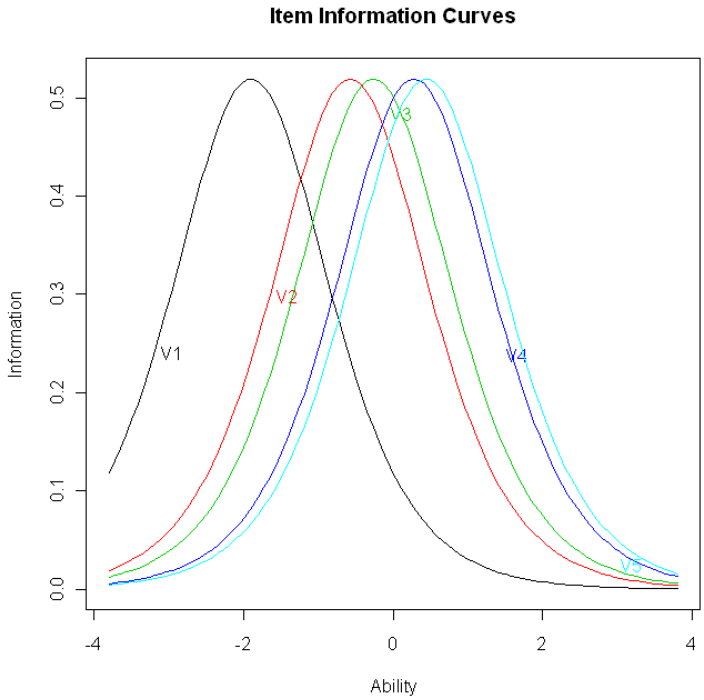


Figure 2.2 Item Information Curves for the 1PL Model

Although the simplicity of the Rasch and 1PL models are theoretically appealing, the presumption that each item discriminates equally well is viewed as a limitation. However, always fitting a Rasch model to the data is recommended, as it provides a baseline to which model fit can be compared to alternative models that incorporate additional parameters (de Ayala, 2009).

2 PL. The 2PL differs from the previous models by allowing unique discrimination values for all items when assessing the probability of a correct item response and is expressed as:

$$P(x_j = 1|\theta, a, b) = \frac{e^{1.7a_j(\theta_i - b_j)}}{1+e^{1.7a_j(\theta_i - b_j)}} \quad (4)$$

Where θ_i is the ability level for person *i;* b_j represents difficulty for item *j*; and a_j represents discrimination for item *j*. Larger values of a_j indicate that the item is better able to differentiate between individuals along the continuum of the latent trait, thus providing more information about the respondents than the Rasch or 1PL models. As can be seen in the ICC plot for the 2PL (Figure 3.1), the five items represented by the colored curves vary in the steepness of their slopes, depicting varying discrimination values that are no longer constrained to a value of one. The ICC plot also reveals that discrimination for Item V5 is somewhat lower compared to the other items, which is illustrated further in the IIC plot (Figure 3.2) by the broad, flat shape of the curve. Conversely, the ICC and IIC for Item V2 indicates a higher discrimination value and suggests that the item provides relatively more information about respondents than the other items, particularly for those with θ between -2 and approximately 1.8.

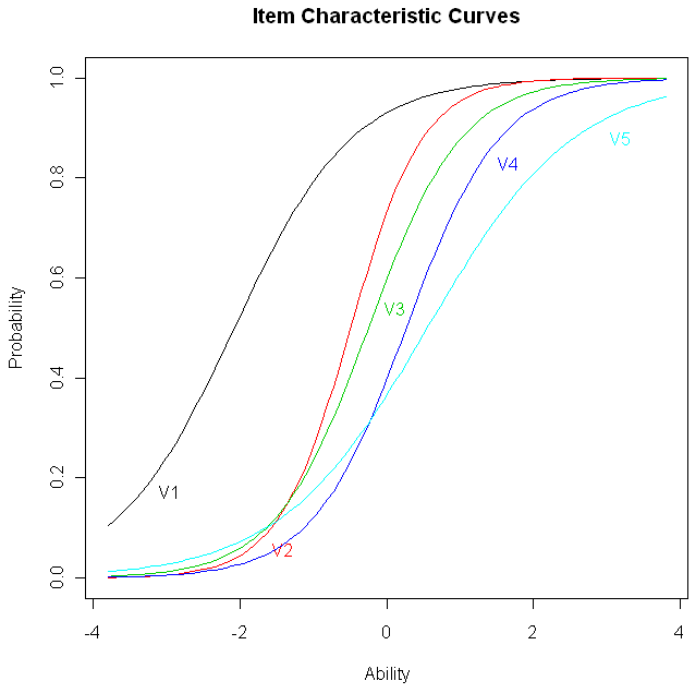


Figure 3.1. Item Characteristic Curves for the 2PL Model

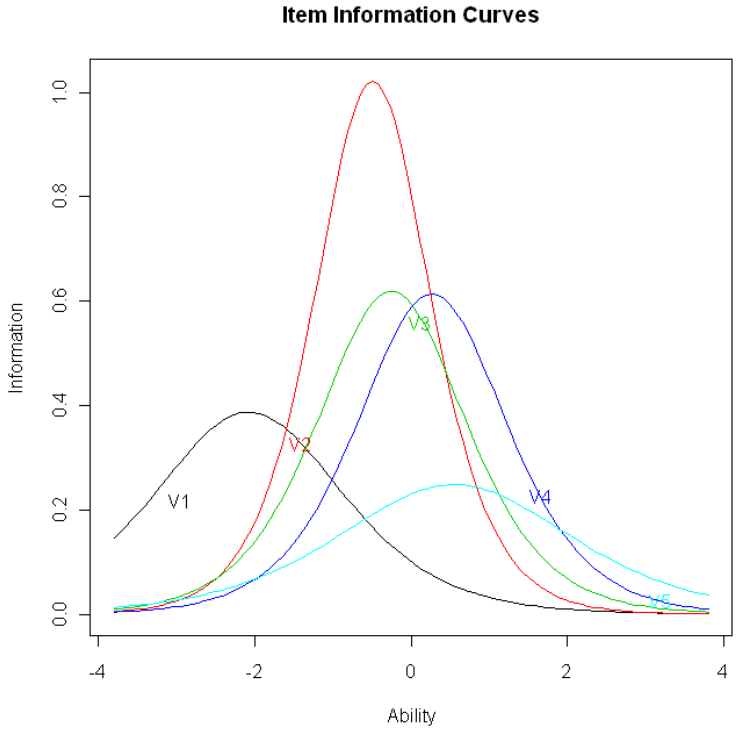


Figure 3.2. Item Information Curves for the 2PL Model

3 PL. The 3PL model is similar to the 2PL model but incorporates an additional parameter for pseudo-guessing. Conceptually, the pseudo-guessing parameter can be envisioned as the probability of a correct response on an item when the respondent has no ability on the trait of interest. Although the 3PL may provide additional information above and beyond the prior models in some situations, it is often not considered relevant on personality assessments and can be problematic. The 3PL is expressed as:

$$P(x_j = 1|\theta, a, b, c) = \frac{c_j + (1 - c_j)e^{1.7a_j(\theta_i - b_j)}}{1 + c_j + (1 - c_j)e^{1.7a_j(\theta_i - b_j)}} \quad (5)$$

Where, θ_i represents the ability level for person i; b_j is difficulty for item j; and c_j is the pseudo-guessing parameter. As shown in the ICC plot for the 3PL model, the addition of the parameter for guessing leads to a flattening out of probability at around 20%, indicating that respondents with lower θ have a higher probability of a correct response than observed with the previous models.

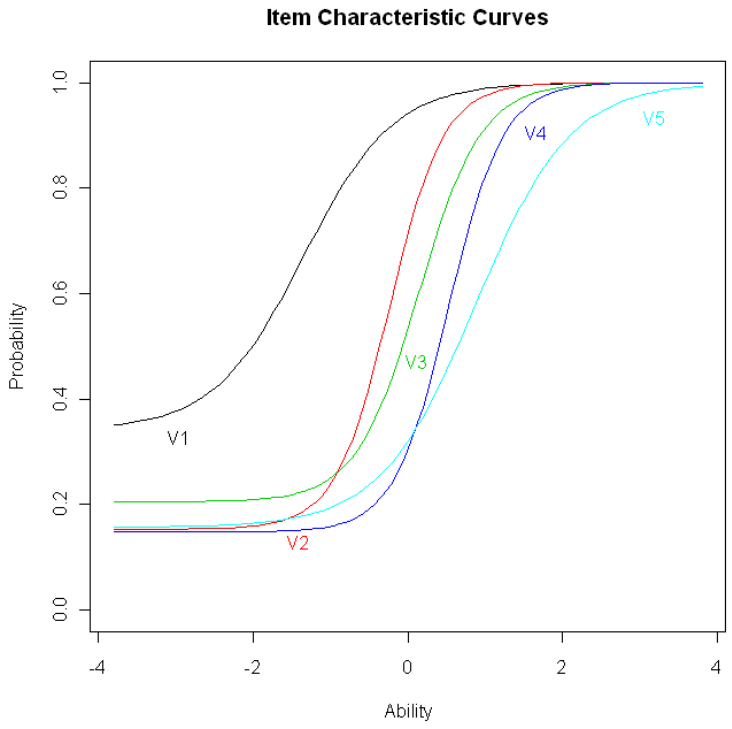


Figure 4.1. Item Characteristic Curves for the 3PL Model

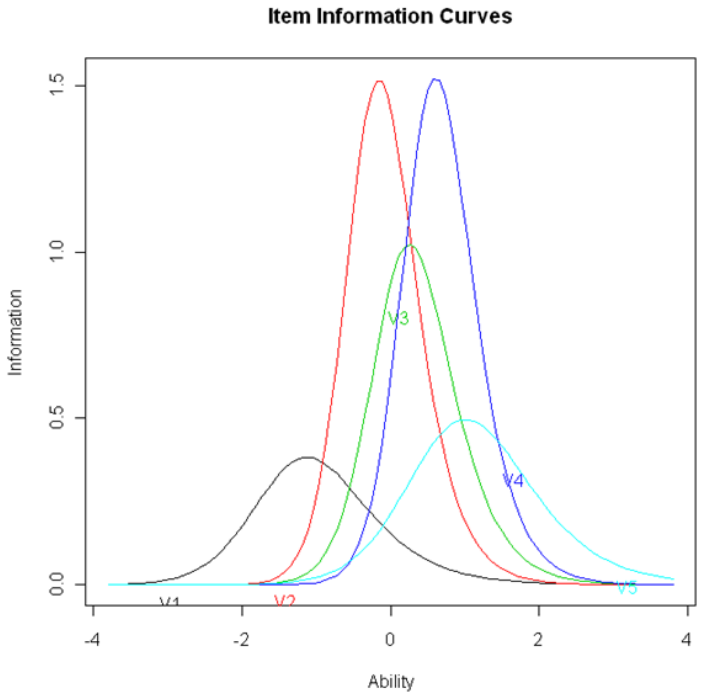


Figure 4.2. Item Information Curves for the 3PL Model

Polytomous IRT Models

Polytomous IRT models are utilized with items that are not scored as correct versus incorrect, as seen with the dichotomous IRT models previously described, but utilize Likert-type scoring where the respondent chooses one among three or more responses indicating their level of agreement with each item. Although θ denotes latent trait levels in polytomous item models the same as in dichotomous item models, item difficulty is referred to as a threshold (δ_{jh}) and denotes the transition point where the probability that a response category is chosen versus another (de Ayala, 2009).

PCM. Corresponding to the Rasch model is the partial credit model (PCM; Masters, 1982), which conceives polytomous items as a serious of dichotomous choices and is expressed as:

$$P(X_j = k|\theta, \delta_{jh}) = \frac{e^{\sum_{h=0}^{X_j}(\theta - \delta_{jh})}}{e^{\sum_{k=0}^{m_j}(\theta - \delta_{jh})}} \quad (7)$$

Where, θ represents person ability, δ_{jh} represents threshold *h* for item *j; k* is the response on item *j*; and m_j represents the maximum possible response categories for item *j*. δ_{jh} can be thought of as the difficulty value for responding with *k* versus *k* – 1. As with the Rasch, item discrimination is constrained to a value of one for the PCM. The number of thresholds for each item is *m*-1, and higher values for δ_{jh} indicate that a respondent is more likely to possess higher levels of the latent trait.

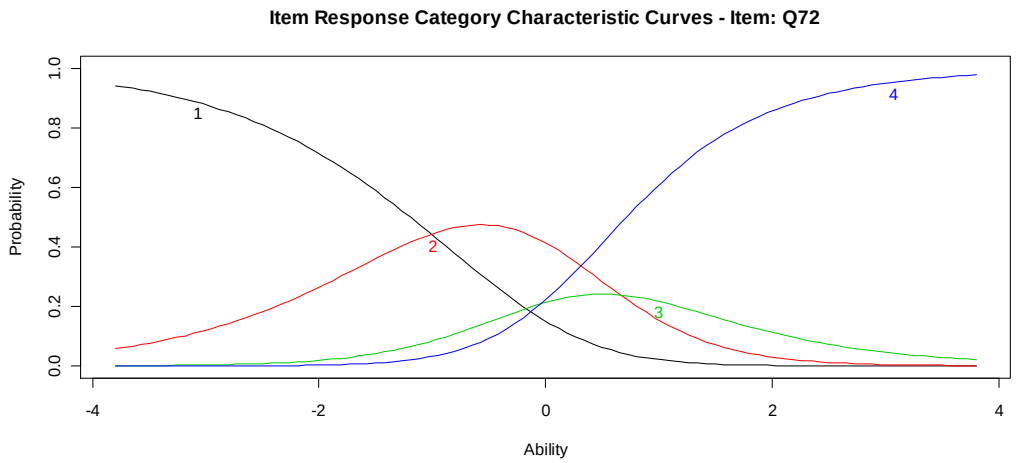


Figure 5.1. Item Response Category Characteristic Curves for the PCM Rasch Model

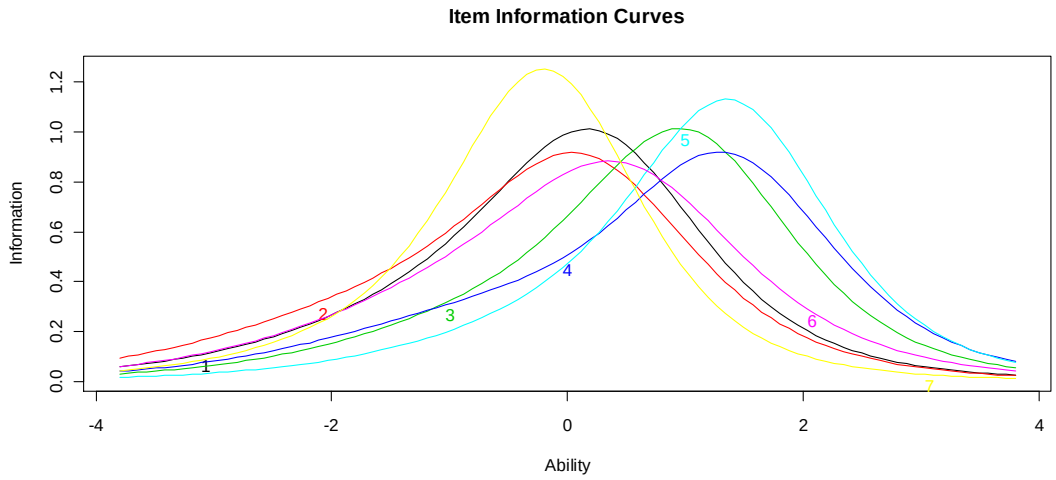


Figure 5.2. Item Information Curves for the PCM Rasch Model

Like their dichotomous counterparts, polytomous IRT models also utilize ICC and ICC plots. As can be seen in the ICC plot (Figure 5.1) of the PCM Rasch model, there are separate curves associated with each individual item, as opposed to the singular curves seen with dichotomous IRT models. These curves depict the various response categories for each item that individuals with a given value of θ are most likely to endorse. The points where the lines of each curve crosses correspond to the threshold. As shown in Figure 5.1, individuals require a θ of about -1.5 before choosing response category 2 becomes more probable than response category 1. Additionally, the plot shows that response category 1 is the most likely response for individuals with θ less than -1.5, while response category 2 is the most probable for a smaller segment of individuals with θ ranging between roughly -1.5 and .5, response category 4 is the most probable for those with θ above .5, and response category 3 is never the most likely to be chosen. IIC plots (Figure 5.2) for polytomous items are interpreted just the same as those for dichotomous items.

GPCM. Corresponding to the 2PL, is the generalized partial credit model (GPCM; Muraki, 1992), which forgoes the requirement that all items have equal discrimination and is expressed as:

$$P(X_j = k|\theta, a_j, \delta_{jh}) = \frac{e^{\sum_{h=0}^{X_j} a_j(\theta-\delta_{jh})}}{e^{\sum_{k=0}^{m_j} a_j(\theta-\delta_{jh})}} \quad (8)$$

Where a_j represents the unique discrimination parameter for each item;

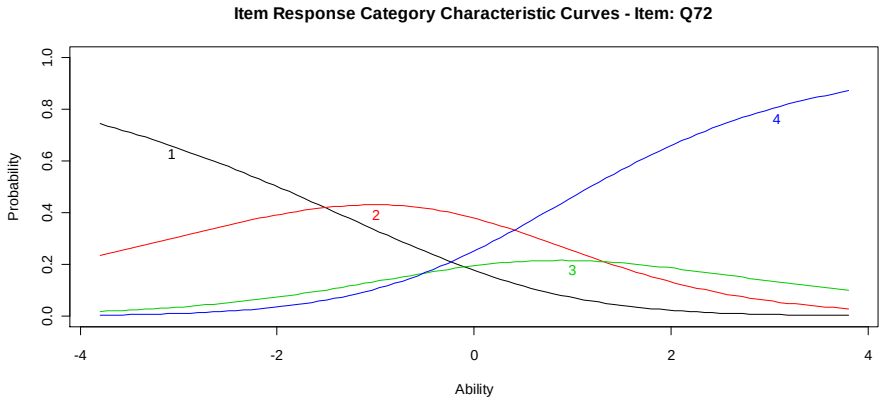


Figure 6.1. Item Response Category Characteristic Curves for the GPCM Model

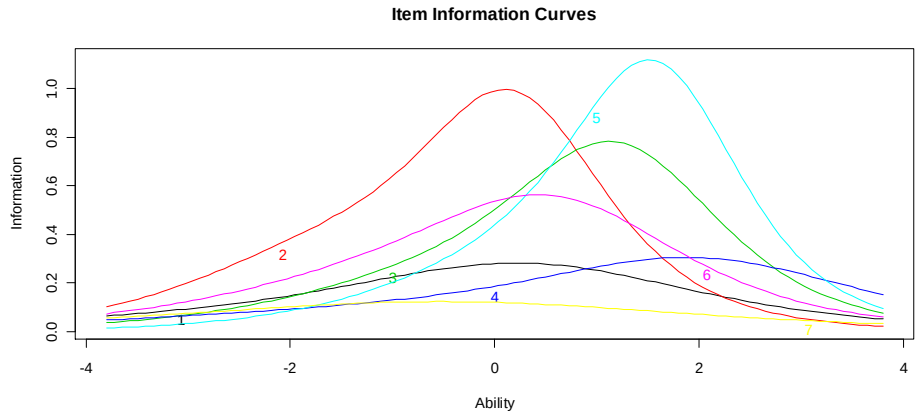


Figure 6.2. Item Information Curves for the GPCM Model

As with the comparison of the ICCs and IICs for the Rasch and 2PL models for dichotomous items, there is more variation seen in the plots for the GPCM as a result of the discrimination parameter no longer being constrained to a value of one. As shown in Figure 6.1 ,

the curves are less steep than that for the PCM Rasch model (Figure 5.1) due to the discrimination value of the item depicted in the example being much lower (.53) when fit with the GPCM model. Additionally, as shown in the IIC plot for the GPCM model (Figure 6.2), there is more variability in the steepness of the peaks of the items to reflect the varying discrimination values in comparison to those previously seen.

GRM. Lastly, the graded response model (GRM; Samejima, 1997) is another popular IRT model for polytomous items that also allows the estimation of unique discrimination values. The formulation of the GRM differs from the models for polytomous items previously described, however, in that the probability of obtaining a given response category is conceptualized as a series of cumulative comparisons rather than dichotomous choices. Mathematically, the GRM is analogous to the 2PL for dichotomous items and is expressed as:

$$P\left(X_j \; or \; higher|\theta, a_j, \delta_{X_j}\right) = \frac{e^{a_j\left(\theta-\delta_{X_j}\right)}}{1+e^{a_j\left(\theta-\delta_{X_j}\right)}} \quad (9)$$

Where θ represents the latent trait; a_j represents discrimination for item *j*; and δ_{X_j} is the threshold between category *k* and *k*-1 for item X_j. ICC (Figure 7.1) and IIC (Figure7.2) plots can also be obtained for the GRM model, and while the ICC plot appears similar to that obtained for the GPCM model (Figure 6.1), the IIC plot depicts curves that are broader, suggesting that the items provide a greater amount of information for a wider span of θ.

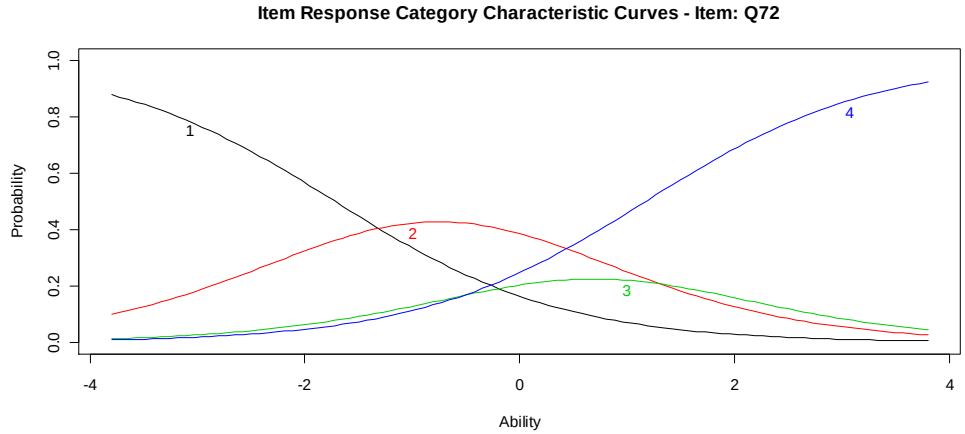


Figure 7.1 Item Response Category Characteristic Curves for the GRM Model

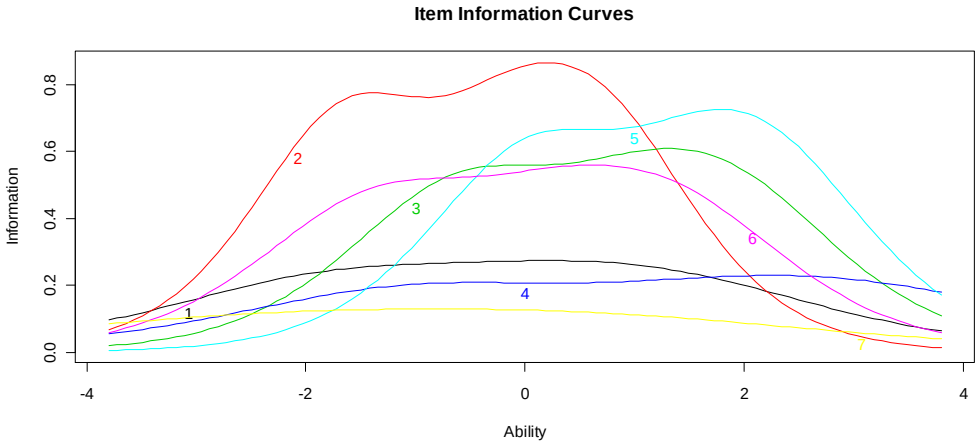


Figure 7.2 Item Information Curves for the GRM Model

Assessing Model Fit

Several methods can be utilized to assess whether a particular IRT model fits the data and to compare model fit between two or more models to determine which fits the data best (Finch & French, 2015). Absolute model fit can be assessed for the test as a whole and each item individually using a chi-square goodness of fit test. This approach tests the null hypothesis that the model fits the data by predicting item responses for each respondent once the model parameters are estimated, then comparing observed response patterns with predicted responses. If observed and predicted response patterns are similar, it can be concluded that the model fits. In

practice, a bootstrap model goodness of fit test is often utilized rather than the standard test, however, as the statistic obtained often diverts from the chi-square distribution. To compare the fit of two models, the relative efficiency—defined as the ratio of information that a more complex model provides (e.g.,3PL) versus a less complex one (e.g.,2PL or 1PL)—can be assessed to determine if additional parameters provide a sufficient increase in information. When the models are nested (e.g.,1PL, 2PL, & 3PL), a difference in the log-likelihood values can also be utilized to compare model fit. When the models are not nested (e.g.,GRM and 2PL), measures of relative fit, the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978), can be compared. Both the AIC and BIC are based upon the log-likelihood values and indicate the amount of variance left unexplained by the model, with more complex models being subject to greater penalization. Consequently, smaller values are indicative of better model fit.

Differential Item Functioning

DIF occurs when an individual's response to an item is associated with group membership (such as gender) that is irrelevant to the construct being measured (Zumbo, 1999). Typically, the group that is thought to have the advantage is referred to as the reference group and the group thought to be disadvantaged is referred to as the focal group, although statistically this designation makes no difference. It is important to note, however, that difference in item response probabilities between groups does not, in-of-itself, indicate DIF, as group differences in the latent trait being measured would be expected to result in group differences in response probabilities for one or more items, a phenomenon termed item impact (Clauser & Mazor, 1998). When respondents are matched on the latent trait—thereby ensuring differences are not due to item impact—and the probability of endorsing an item differs across groups, then DIF is present.

This indicates that the item is performing differently for different groups. Test equity across groups is an essential component of instruments if we hope to draw accurate conclusions about the population being measured. Therefore, determining whether DIF is present within the items of an instrument is a crucial component to establishing validity.

Types of DIF

There are two general categories of DIF—uniform and nonuniform DIF (Zumbo, 1999). Uniform DIF is present when the probability of endorsing an item is uniformly higher for the reference or focal group across all levels of θ. As seen in Figure 8.1 the ICC for the reference and focal groups are parallel, indicating no interaction between ability and group membership. In the case of nonuniform DIF, there *is* an interaction between ability and group membership such that the probability of endorsing an item for the two groups varies across the continuum of θ. As shown in Figure 8.2, the probability of endorsing the item is lower for Group 2 (represented by the dashed line) at levels of θ below average, but higher at levels of θ that are average and slightly higher, resulting in the crossing of ICC's. Although nonuniform DIF is said to be less common, its detection is no less important than uniform DIF (Finch & French, 2007).

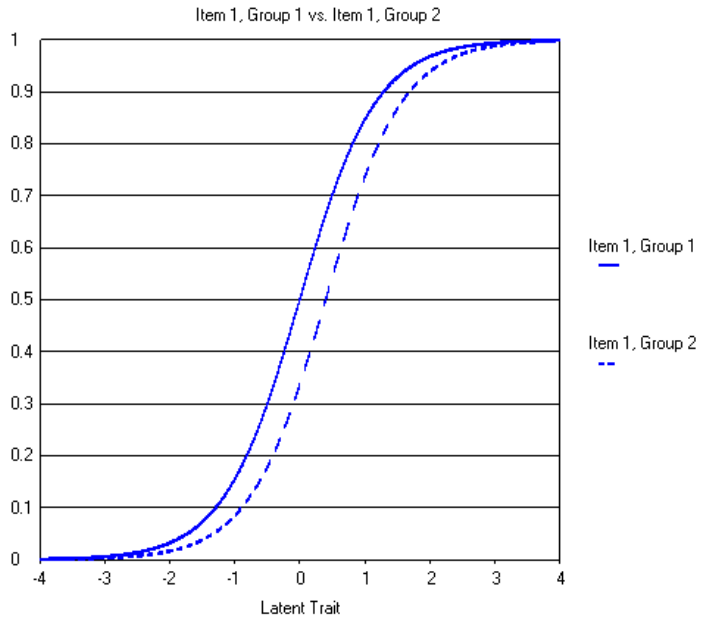


Figure 8.1. Uniform DIF

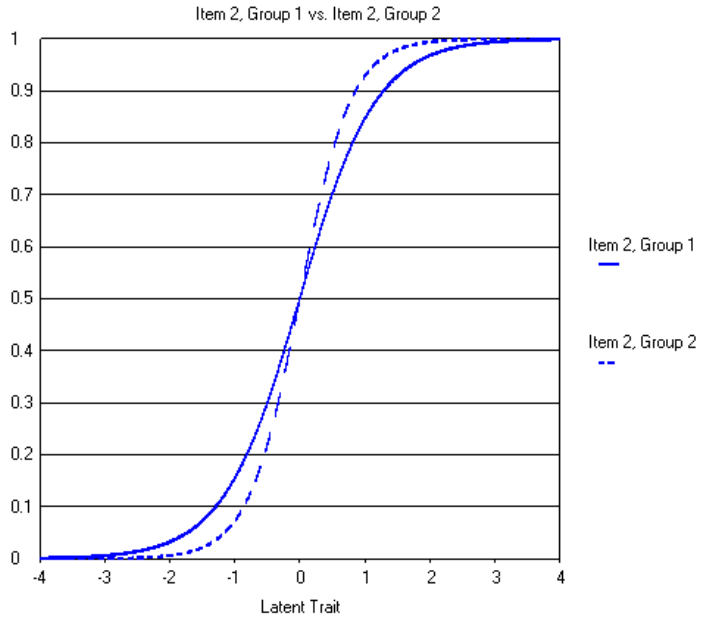


Figure 8.2. Nonuniform DIF

DIF Detection Methods

Currently, there are several methods available for the detection of DIF, including the Mantel-Haenszel (MH; Holland & Thayer, 1998), simultaneous item bias test (SIBTEST; Shealy

& Stout, 1993), logistic regression (LR; Swaminathan & Rogers, 1990), IRT likelihood ratio test (IRT LR; Thissen, Steinberg, & Wainer, 1998), and the DIF detection method selected for use in the proposed study, multiple indicators multiple causes (MIMIC; Jöreskog & Goldberger, 1975). Each of these approaches demonstrate strengths in detecting DIF given certain circumstances and will be briefly described in the following paragraphs.

IRT LR. The IRT LR statistic consists of comparing the fit of two IRT models using the likelihood ratio test statistic (Thissen et al., 1998). The comparison assesses whether there is a significant difference in the model fit after constraining an item to have the same location across the reference and focal groups versus when the item is allowed to differ in its location. To compare the two models, a log-likelihood statistic (LL_{equal}) if first calculated for the constrained model and is expressed as:

$$LL_{equal} = \sum_{G=1}^{2}\sum_{p=1}^{N} 1n\left[\sum_{1}^{q} \prod_{i=1}^{n\ items} \left(T_{iG}\left(u_{ip}G\right)\phi_{G}\left(\theta\right)d\theta\right)\right], \qquad (10)$$

Where $T_{iG}\left(u_{ip}\,G\right)$ is the ICC parameters for group G, equally constrained for reference and focal groups; and $\phi_i\,G\,(\theta)$ is the distribution of the latent trait for group *G.* A second IRT model is then fit to the data, with the parameter being examined for DIF allowed to differ and a second log likelihood then calculated ($LL_{unequal}$) and the two log likelihoods are calculated. If DIF is not present, then the two location estimates should be the same when the item location can vary across the groups. IRT LR has proven to be an effective method of DIF detection in a variety of circumstances. However one of the limitations is that it requires a large sample size.

Mantel-Haenszel. The MH statistic (Holland & Thayer, 1988) is a widely used and computationally simple nonparametric DIF procedure. To estimate DIF, the MH first arranges item responses for the reference and focal groups into a 2 x 2 contingency table and compares

the probabilities of a correct response for respondents who have been matched on ability. Separate tables for each test item are constructed depicting Group x Item responses at each score level. A chi-square statistic is then calculated for each of the tables testing group membership (i.e., focal vs reference), and item response and is expressed as:

$$\chi^2_{MH} = \frac{\left[\left|\sum_{j=0}^{J} A_j - \sum_{j=0}^{J} E(A_j)\right| - 0.5\right]^2}{\sum_{j=0}^{J} var(A_j)} \quad (11)$$

Where A_j is the number of reference group respondents for total score *j* who answered the item correctly; $E(A_j)$ is the expected number of reference group respondents for total score *j* who answered the item correctly if no DIF is present.

$$E(A_j) = \frac{N_{Rj} N_{c.j}}{N_{..j}} \quad (12)$$

and:

$$var(A_j) = \frac{N_{Rj} N_{Fj} N_{c.j} N_{w.j}}{N^2_{..j}(N_{..j} - 1)} \quad (13)$$

Although the MH is widely used for a range of purposes, including DIF detection, the procedure only assesses uniform DIF.

SIBTEST. SIBTEST (Shealy & Stout, 1993), is another nonparametric approach that has been shown to be useful in detecting uniform and nonuniform DIF. SIBTEST procedure is capable of detecting DIF at the item level or bias as a characteristic of the test as a whole, emphasizes matching the reference and focal group more accurately than many other procedures, and is based on the assumption that DIF occurs as a result of group differences on a secondary dimension present within the items. DIF is estimated by first creating two subsets of responses from the reference and focal groups—one where items are suspected to exhibit DIF and another purported to exhibit no DIF. The two groups are then matched on scores for the subset thought to

be free of DIF. *B* can then be interpreted as the average proportion correct for respondents on the subset suspected of exhibiting DIF. The formula for SIBTEST can be expressed as:

$$\widehat{\beta} = \sum_{k=0}^{n} \hat{p}_x \left(\hat{P}_R [T_x] - \hat{P}_F [T_x]\right) \quad (14)$$

Where $\hat{p}_x$ is the proportion of subjects with a matching subtest score of $X = x$; $\hat{P}_R [T_x]$ represents the proportion of subjects in the reference group with a matching subset score of *x* answering the item correctly; and $\hat{P}_F [T_x]$ is the proportion of subjects in the focal group with a matching subset score of *x* answering the item correctly.

Logistic Regression. Like the MH procedure, LR **(**Swaminathan & Rogers, 1990) is also widely used in a variety of purposes beyond DIF detection and can be used to detect both uniform and non-uniform DIF. LR in the context of DIF works by predicting the probability of a correct item response as a function of total score, group membership, and the interaction between group membership and total score. The LR model can be expressed as:

$$P\ (u = 1) = \frac{e^z}{1 + e^z} \quad (15)$$

Where $z = \beta_0 + \beta_1 X + \beta_2 G$ assesses uniform DIF, while $z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$ assesses nonuniform DIF; $P\ (u = 1)$ is the probability of a correct item response, *X* is the total test score, and *G* represents group membership (the reference group is denoted as *G* = 1 and the focal group is denoted as *G* = 0). A statistically significant interaction between group membership and ability ($\beta_3 \neq 0$)indicates the presence of nonuniform DIF, while a significant effect ($\beta_2 \neq 0$) suggests uniform DIF.

MIMIC. The MIMIC model (Jöreskog & Goldberger, 1975; Muthén, 1989) is also known as confirmatory factor analysis (CFA) with covariates. Uniform DIF detection with the MIMIC method occurs through the estimation of direct and indirect effects for a grouping

variable (e.g., gender). With MIMIC, a model is fit in which the latent trait is measured by the item and, as shown in Figure 9, the indirect effect regresses the latent trait onto the grouping variable and determines whether mean differences on the latent trait exist, while the direct effect path regresses the item response onto the grouping variable. After controlling for the mean group differences on the latent trait (i.e., impact), the difference is the estimation of DIF present in the item.

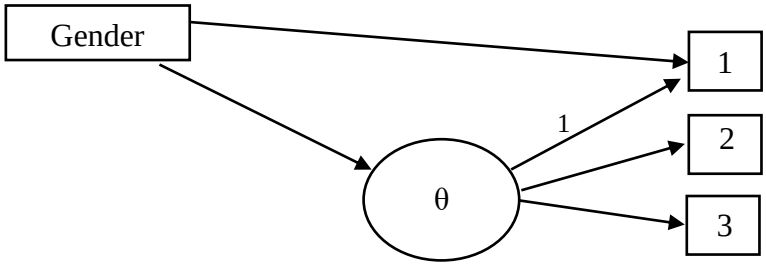


Figure 9. MIMIC Model

The MIMIC model when applied to DIF can be expressed as:

$$y_i^* = \lambda_i \eta + \beta_i z_k + \varepsilon_i \quad (16)$$

Where y_i^* represents the latent response underlying the item response for item *i;* λ_i is the factor loading (discrimination parameter); η is the latent trait; z_k represents a dummy variable indicating group membership; β_i is the slope for the group variable and item response; and ε_i represents random error. Although the MIMIC model is a case of a confirmatory factor analysis model with covariates, the relationship between IRT models and confirmatory factor analysis has been well established and both are capable of DIF detection, with the parameter estimates obtained using the MIMIC model being easily converted to IRT parameter estimates (e.g., ability and item discrimination) common to IRT models (Finch, 2005; Jones & Gallo, 2002; MacIntosh & Hashim, 2003). The MIMIC model has been found to perform well in detecting uniform and

nonuniform DIF in both dichotomous and polytomous items (Balut & Suh, 2017; Chun, 2014; Finch, 2005; Lee, Bulut, & Suh, 2017; Woods, 2009).

Often, DIF is evaluated in tests comprised of one latent trait, known as unidimensional. Traditionally, many DIF approaches in fact assume unidimensionality of the test when assessing items for DIF (de Ayala, 2009). However, many tests—especially those that measure complex constructs—are intentionally designed to assess multiple latent traits, known as multidimensional (Ackerman, 2005). In such cases, the DIF approach must be able to factor in the presence of more than one latent trait to avoid falsely-positive errors in DIF detection (Mazor, Hambleton, & Clauser, 1998). As previously noted, the SOI-R assesses sociosexuality as a multidimensional construct consisting of three latent traits—behavior, attitude, and desire, so it is therefore necessary to utilize a methodology that can detect DIF in multidimensional instruments. Currently, many DIF methods have been adapted from their more traditional counterparts that were designed for DIF detection in unidimensional instruments, including LR, IRT LR, and the MIMIC model (Lee et al., 2016; Mazor et al., 1998; Suh & Cho, 2014). However, because the findings from previous research indicate that demographics such as race, education level, and age also influence both sociosexuality and endorsement of SDS (which is suspected to underly response bias on the items of the SOI-R), having the ability to investigate more than one demographic variable is another chief concern for the proposed study (Allison & Risman, 2013; Penke & Asendorpf, 2008). Given that a primary advantage of the MIMIC model is that it also allows for a comprehensive examination of the relationship between multiple background variables and the latent trait and would therefore be able to account for multiple demographic categories, it was selected as the DIF detection method for this study (Chun, 2014; Lee et al. , 2016; Muthén, 1988; Teresi, 2006). The MIMIC model has also been shown to perform well in

detecting DIF in multidimensional tests—particularly those with fewer items— and had lower Type 1 error rates than LR (Bulu & Suh, 2017; Lee et al., 2016). As such, the MIMIC model appears to be an appropriate candidate for DIF detection in the SOI-R.

Hypothesis and Research Questions

Due to the evidence throughout the literature indicating that endorsement of the SDS is still prevalent in modern culture and has been shown to influence responses to items assessing sexual attitudes and behaviors (particularly among women), it was expected the results of this study would detect DIF in the items of the SOI-R (Fenton et al., 2001; Mitchell et al., 2018; Peterson & Hyde, 2010; Streger et al.,2013). More specifically, it was expected that items 1 through 3 (Behavior) would exhibit DIF favoring men given the findings throughout the literature suggesting men and women use different strategies when reporting past partners and are prone to false accommodation of responses (Mitchell et al., 2018). The occurrence of gendered DIF favoring men would suggest that men have a higher probability of endorsing an item than women, even after the instrument assesses both genders as having equally unrestricted orientations.

It was also predicted that items 4 and 5 (Attitude) would exhibit DIF favoring men due to women accommodating their responses to be in line with gendered expectations that women hold favorable attitudes towards sex only when in the context of a committed relationship (Peterson & Hyde, 2010). The third item of the attitude subscale (Item 6), "*I do not want to have sex with a person until I am sure that we will have a long-term, serious relationship*" was also expected to exhibit DIF favoring men after reverse scoring of the item. There were no specific predictions regarding Items 7 through 9 (Desire); however, these items were also examined for DIF.

CHAPTER 3: METHODOLOGY

Participants

The goal of the present study was to assess whether the items comprising the SOI-R exhibited DIF for gender. The data for which the analyses was performed consisted of responses to the SOI-R provided by 1970 individuals (1022 women and 948 men) who had taken part in a prior online study. The main purpose of the previous study was to examine interest in group sex; however, the survey also contained a variety of other personality and attitude measures, including the SOI-R.

The study was first approved by the Ball State Institutional Review Board and participants were recruited through a solicitation notice that was posted on various online psychology research forums, social media websites (e.g., Twitter, Reddit, & Facebook), and the volunteer section of Craigslist. The solicitation notice was intentionally vague and simply asked respondents to participate in a study examining attitudes towards and experience with a variety of sexual behaviors. All participants were required to attest that they were at least 18 years of age to take part in the study. No identifying information was collected and participants were ensured that their responses would remain anonymous.

Prior to cleaning data, 3,127 individuals provided informed consent between November of 2016 and 2018. To minimize the influence of potentially confounding variables, only data for respondents who reported their birth sex as male or female and whose current gender identity matched their birth sex were analyzed. Consequently, the data for five respondents who identified their birth sex as "*other*" and 110 who identified their current gender identity as being anything other than male or female (e.g., genderfluid, bigender, or transgender) were removed. An additional 281 responses were removed due to failure to indicate birth sex and/or gender identity. Another 15 individuals indicated that they were born male but currently identified as

female, while 11 indicated the inverse and were therefore not included in analyses, leaving 2705 responses. However, 699 participants (367 women and 332 men) did not provide responses to any items of the SOI-R. Listwise deletion was used for the removal of any remaining responses where at least 1 item of the SOI-R was left blank, which removed another 12 (7 women and 5 men) responses. Finally, although the data for various sexual orientations were analyzed, responses (n = 31) for individuals who identified as asexual (17 women and 6 men) were removed prior to analyses.

As previously noted, the SOI-R was included among several other survey items that assessed participants' attitudes towards and experiences with group-sex and other sexual behaviors, personality traits, and relationship satisfaction. However, given that the focus of this study was to examine the items of the SOI-R for DIF, participants' responses on other assessments (aside from basic demographic information) are not reported in the results.

Participants' demographic information collected included race, highest education level, relationship status, current student status, age, and Kinsey scale scores— which assesses respondents' degree of heterosexuality/homosexuality on a continuum . Race was assessed by having participants choose from the options of, "*African American or Black*," "*Asian or Pacific Islander*," "*White or European American*," "*Hispanic, Native American or Alaskan*," "*Biracial or Multiracial*," and "*Other*". Highest education level had options ranging from "*did not complete high school*," to "*doctoral or advanced professional*." Relationship status was assessed with an item inquiring whether participants were currently in a relationship, with "*yes*," or "*no*" response options. Current student status was assessed with an item that read, "*Are you currently a full or part-time college student*?" Response options included "*yes*" and "*no*." Age was assessed in years and Kinsey scale scores were assessed on a 6-point scale, where *0* indicated

"*completely heterosexual*," and *6* indicated "*completely homosexual*." Although the scale included a 7th point indicating asexuality, as previously described, this was removed prior to analysis.

Chi-square tests of association were performed on categorial demographic variables to determine whether men and women differed significantly on any of the categories. Variables with three or more categories were collapsed into two categories for comparison purposes. As a result, race was collapsed into "*white*" versus "*non-white*," and education level was collapsed into "*at least some college*" versus "*no college*." Kinsey scale scores and age were compared with independent samples *t*-tests. Demographic variables that were thought to potentially also be associated with DIF, such as Kinsey scale scores and race, were included in the MIMIC model as grouping covariates.

Materials

The instrument that was the focus of this study was the 9-item SOI-R. Item responses are summed for each of the three latent scores and then can be summed altogether for a global sociosexuality score. A low score is indicative of a restricted orientation, whereas a high score is indicative of an unrestricted orientation. Put another way, people with lower scores have a preference for only engaging in sex with people in which emotional intimacy has been established, whereas those with higher scores have no such restrictions to engaging in sex with others. Items 1 through 3 assess sociosexual behavior and were assessed on a 9-point scale. Response choice "*1*" represented having had no sexual partners; "*2*" represented one partner; "*3*" two partners in the past year; "*4*" three partners; "*5*" four partners; response "*6*" represented five to six partners; "*7*" was seven to nine; "*8*" ten to nineteen partners; and "*9*" indicated having had 20 or more sexual partners. Items 4 through 6 assessed sociosexual attitude on a 9-point scale where response choices ranged from *1* being "*strongly disagree*" to *9* representing

"*strongly agree*." Items 7 through 9 assessed sociosexual desire on a 9-point scale with response options ranging from *1* representing "*never*" to 9 indicating, "*at least once a day*." All items of the SOI-R are presented in full in Table 9.

Although IRT methods of obtaining information about the reliability of instrument have been shown to be more accurate and informative than classical test methods, Cronbach's alpha for the SOI-R was also obtained for the total sample and by gender. A confirmatory factor analysis (CFA) was also performed to ensure that the 3-factor solution of the SOI-R was an adequate fit for the data.

Procedure

Multidimensional IRT (MIRT) analyses. Prior to assessing the items for DIF, a MIRT analysis was performed on the data to determine item difficulty and discrimination values. Given that the items of the SOI-R were scored on a 9-point scale, item difficulty is represented as the values of eight thresholds, which reflect the level of the latent trait needed for respondents to have at least a 50% probability of endorsing the next response choice. Item thresholds and discrimination parameters provided by the model that best fit the data were then used in the interpretation of DIF to determine the nature of subgroup differences (e.g., whether one group had a higher probability of endorsing a response choice of a particular item at a lower level of θ). Three separate MIRT models were fit to the data using the MIRT package in R (Chalmers, 2012) and were then compared to determine which model best fit the data. These included the PCM Rasch (Masters, 1982), the GPCM (Muraki, 1992), and the GRM (Samejima, 1969). A multidimensional 1PL was not available, so this model was not fit to the data. To determine best model fit, several methods were utilized. These included examining the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) for each model and comparing them between models to assess which model had the lowest of these values, with lower values

indicating better model fit. Nested models (e.g., Rasch and GPCM) were also compared using the likelihood ratio test, which signifies whether one model fit better than the other.

DIF analyses. Uniform and non-uniform DIF were assessed simultaneously with MIMIC models using Mplus software version 7.11 (Muthen & Muthen, 2013). A CFA model with three factors with covariates was specified, with Items 1 through 3 belonging to Factor 1: Behavior, Items 4 through 6 belonging to Factor 2: Attitude, and Items 7 through 9 belonging to Factor 3: Desire. Within the same model, Factor 1: Behavior was then regressed on Gender, (dummy coded as Men = 0, Women = 1). To control for the effects of additional demographic variables, the factor was also regressed on Race (White = 1, Non-White = 0) and Kinsey scale scores. The item was then regressed on Gender, which served as the significance test for uniform DIF for each item. An interaction variable consisting of Gender and the latent trait was also regressed onto the item, which served as the significance test for non-uniform DIF. The two remaining items in each factor that were not being tested for DIF where used as anchor items. This process was repeated for each individual item, with each item related to its specific factor. Because there were so few items in each factor and many items were found to exhibit DIF, item purification, which would have removed items flagged for DIF and reran the process again until all items with DIF had been removed, was not performed. Ideally, this process would have produced a set of anchor items that were free of DIF, therefore reducing the likelihood that subsequent items in the scale would falsely test positive for DIF (Wang et al., 2009). Alternatively, scale items that are thought to be DIF free can be selected a priori as anchor items from which remaining DIF suspected items are tested. Because the literature suggested that it was plausible that all items of the SOI-R could contain DIF, specific anchor items were not selected a priori. Instead, for each

item that was examined for DIF individually with the MIMIC model, the two remaining items in each three-item factor which were not being tested in that model were used as anchor items.

CHAPTER 4: RESULTS

Participant Demographics

The final data for this study consisted of 1970 individuals (1022 women and 948 men) who provided responses to items of the SOI-R. Demographic results are presented in Table 1 according to gender.

Table 1. Participant Demographics

	Age	Kinsey Scale Scores	Currently in a Relationship	White Participants	Current Students	At Least Some College
	M (SD)	*M (SD)*	*% / n*	*% / n*	*% / n*	*% / n*
Men $n = 947$	36.61 (14.95)	2.37 (1.81)	66% $n = 626$	81.9% $n = 776$	33.3% $n = 315$	90% $n = 853$
Women $n = 1022$	27.67 (9.82)	2.28 (1.44)	67.2% $n = 687$	69.1% $n = 706$	57.3% $n = 585$	87.6% $n = 853$

Note. Kinsey scores represent means based on a 6-point scale, with lower scores indicating orientations that were more heterosexual.The category of "white participants" and "at least some college" reflect collapsed categories for race and education level.

Chi-square tests of association determined that the two subsamples differed significantly on Race, $\chi^2(1) = 43.08$, $p < .001$, with men having a significantly greater proportion of participants who identified as white. Men and women also differed on Current Student Status, $\chi^2(1) = 113.94$, $p < .001$, with significantly more women reporting they were currently students, and Age, $t(1620.41) = 15.54$, $p < .001$, with the average age of men being higher than for women (Demographics reported in Table 1). No significant differences were found between the demographic variables of Education Level, $\chi^2(1) = 2.85$, $p = .092$; Relationship Status, $\chi^2(1) = .312$, $p = .576$; or Kinsey scale scores, $t(1790.9) = 1.24$, $p = .214$. Full demographic proportions for Race and Education prior to collapsing categories are reported in the appendix .

Reliability. Using a 95% confidence interval, the results revealed that the SOI-R demonstrated good reliability with Cronbach's alpha of .86, with confidence intervals ranging from .85 to .87. Item correlation statistics showed strong correlations between each item and the rest of the scale that ranged from .52 (Item 1) to .75(Item 5), with the majority of items falling around an *r* of .66. Using a 95% confidence interval, Cronbach's alpha for women's responses was .86 [.85, .88], and slightly lower for men's, at .84 [.82, .85].

Factor Analysis. A confirmatory factor analysis (CFA) was performed on the SOI-R to determine how well the three-factor solution fit the data. The root mean square error of approximation (RMSEA) was .076, with a 90% confidence interval of 0.069 to 0.084, suggesting adequate fit (SRMSR = .061). Values for the Comparative Fit Index (CFI; .97) and the Tucker-Lewis Index (TLI; .96) also suggested that the three-factor structure of the SOI-R was an acceptable fit for the data. Results revealed that the three factors were moderately correlated with one another. Factor 1: Behavior and Factor 2: Attitude had a correlation of .45 ($p < .001$); Factor 1: Behavior and Factor 3: Desire had a correlation of .32 ($p < .001$); and Factor 2: Attitude and Factor 3: Desire, a correlation of .49 ($p < .001$).

Table 2. CFA Estimate for the 3-Factor Solution of the SOI-R

	Item	Unstandardized Estimate	Standard Error	Standardized Estimate
	Item 1	1.00		.49
Factor 1: Behavior	Item 2	2.24	0.10	.85
	Item 3	2.83	0.13	.97
	Item 4	1.00		.73
Factor 2: Attitude	Item 5	1.45	0.04	.90
	Item 6	1.08	0.04	.74
	Item 7	1.00		.83
Factor 3: Desire	Item 8	0.93	0.02	.81
	Item 9	1.00	0.03	.85

Note. Factor loadings are represented as the standardized estimates. The first item in each factor has been constrained to a value of 1 and serves as the reference.

Item Descriptive Statistics. Means and standard deviations for each of the items by gender were also obtained and are presented in Table 3. These statistics were also obtained for the three subscales of the SOI-R and are reported in Table 4.

Table 3. Descriptives for SOI- Items for Men and Women.

	Men		Women	
	M	SD	M	SD
Item 1	2.83	1.93	2.92	1.88
Item 2	3.28	2.56	3.01	2.34
Item 3	3.75	2.77	3.57	2.64
Item 4	7.07	2.45	6.76	2.68
Item 5	6.38	2.75	5.22	3.12
Item 6 (R)	6.83	2.52	6.07	2.86
Item 7	6.19	2.39	4.55	2.47
Item 8	4.70	2.45	3.66	2.31
Item 9	4.80	2.52	3.21	2.21

Note. Means reflect reverse scoring for Item 6 (R)

Table 4. Descriptives for Men and Women on the SOI-R Subscales and Scale

	Behavior	Attitude	Desire	Total
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Men n = 947	3.28 (2.07)	6.76 (2.20)	5.23 (2.14)	5.09 (1.64)
Women n = 1022	3.16 (1.96)	6.02 (2.52)	3.81 (2.07)	4.33 (1.75)

Note. Scores reflect the means for the sum of items in each subscale (9-pt scaling). Item 6 was reversed scored prior to analyses.

MIRT Results for Total Sample. Prior to performing analyses for DIF detection, three MIRT models were fit to the data for all participants (men and women combined) using the MIRT package (Chalmers, 2012) for R statistical software. These models included multidimensional versions of the Rasch, the GPCM 2PL, and the GRM. A likelihood ratio test was used to compare the nested models (i.e. GPCM and Rasch), with the results indicating that the fit of the two models was not the same, χ^2 (6) = 2736.13. As shown in Table 5, fit indices were smaller for the GPCM 2PL than for the Rasch suggesting that the GPCM fit the data better.

Table 5. Indices for MIRT Analyses of the SOI-R Items for Total Sample

	AIC	AICc	SABIC	HQ	BIC	Log L.
Rasch	63485.39	63491.53	63664.62	63638.95	63902.90	-31667.69
GPCM	60761.26	60768.44	60954.84	60927.11	61212.17	-30299.63
GRM	60191.38	60198.55	60384.95	60357.23	60642.29	-30014.69

Note. Smaller values of indices indicate better fit.

Although the GPCM and GRM could not be compared via the likelihood ratio test due to the fact that the models are not nested, the relative fit indices (i.e., AIC, BIC, AICc, SABIC, and HQ) were smaller for the GRM suggesting that the GRM was the better fitting model. MIRT analyses were then performed for the data of men and women separately. Results are reported in the following paragraphs.

MIRT Results for Men. As with analysis for the total sample, the results of the MIRT analyses for men revealed that the GRM provided the best fit to the data. A log likelihood ratio test comparing the GPCM and Rasch determined that the fit of the two models differed significantly, χ^2 (6) = 1239.75, $p < .01$. An examination of the indices (Table 6) revealed that the GPCM was a better fit than the Rasch, and although the fit of the GRM could not be compared statistically to that of the GPCM, the smaller relative fit indices suggested it fit the data best.

Table 6. Indices for MIRT Analysis of SOI-R Items for Men

	AIC	AICc	SABIC	HQ	BIC	Log L.
Rasch	30495.65	30508.95	30620.34	30634.03	30858.53	-15172.83
GPCM	29267.90	29283.51	29402.56	29417.35	29659.81	-14552.95
GRM	28966.88	28982.49	29101.55	29116.34	29358.80	-14402.44

Note. Smaller values of indices indicate better fit.

IRT Results for Women. The results of the MIRT analyses for women also revealed that the GRM fit the data best. The results of a log-likelihood ratio test indicated that the fit of the GPCM and Rasch differed significantly (χ^2 (6) = 1495.335, $p < .01$), with the indices (Table 7)

suggesting that the GPCM was a better fit. Because indices were again smallest for the GRM, it was determined to be the best fit of the three models.

Table 7. Indices for MIRT Analyses of SOI-R Items for Women.

	AIC	AICc	SABIC	HQ	BIC	Log L.
Rasch	32654.62	32666.96	32784.50	32794.52	33022.70	-16252.31
GPCM	31171.26	31185.74	31311.53	31322.35	31568.79	-15504.63
GRM	30932.46	30946.93	31072.72	31083.55	31329.99	-15385.23

Note. Smaller values of indices indicate better fit.

Taken together, these results indicate that the GRM was the best fit for the data. As previously noted in Chapter 2, the GRM provides both a unique discrimination value for each item, reflecting how well the item differentiates between respondents at varying levels of the trait, and *m-1* thresholds (where *m* depicts the total number of response options for item), reflecting the amount of θ needed for a 50% or greater probability of endorsing a particular response or higher versus lower response options. Because the SOI-R was scored using a 9-point format, there were eight thresholds per item (denoted as b_1 to b_8 in Table 8). These thresholds and discrimination parameters (denoted as *a* in Table 8) for both men and women were obtained using the GRM and are reported in Table 8.

Overall, these results indicated that all nine items of the SOI-R were good discriminators, suggesting that all items were adept at differentiating between men and women with restricted sociosexual orientations versus unrestricted orientations. The results further revealed that items 2 and 3— which both assessed past numbers of sexual partners outside the context of a long-term relationship—were the best discriminators for both men and women, with both having values above 4.0. Item 5, which assessed anticipated comfortability and enjoyment of casual sex with different partners, was also revealed to be an excellent discriminator. Item 1 was shown to have the lowest discrimination value for both men and women at slightly over 1. Item parameters for

men and women are discussed in greater detail for each item individually in the following paragraphs, in addition to MIMIC results for DIF. Plots depicting response categories for each item and total information provided are presented in the appendix.

Table 8. GRM Item Parameters for Men and Women's Responses on the SOI-R.

		a	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
Item 1	*W*	*1.284*	*-5.355*	*-3.973*	*-3.035*	*-2.305*	*-1.782*	*-1.315*	*-0.591*	*2.069*
	M	1.043	-4.635	-3.579	-2.996	-2.26	-1.874	-1.319	-0.764	1.647
Item 2	*W*	*4.434*	*-8.775*	*-7.133*	*-5.84*	*-4.451*	*-3.555*	*-2.718*	*-1.115*	*1.417*
	M	4.01	-7.889	-5.719	-4.597	-3.675	-3.021	-1.956	-0.756	1.507
Item 3	*W*	*4.337*	*-7.448*	*-5.464*	*-4.384*	*-3.327*	*-2.284*	*-1.193*	*0.055*	*1.887*
	M	4.172	-7.094	-4.979	-4.076	-2.857	-2.051	-1.067	0.064	1.826
Item 4	*W*	*2.451*	*-0.535*	*0.248*	*1.073*	*1.703*	*2.489*	*2.962*	*3.63*	*4.239*
	M	3.064	-0.407	0.63	1.767	2.429	3.488	4.051	4.68	5.144
Item 5	*W*	*4.024*	*-2.989*	*-1.747*	*-0.652*	*0.178*	*0.78*	*1.427*	*2.576*	*3.675*
	M	3.669	-1.467	-0.282	0.95	2.084	2.86	3.307	4.14	5.148
Item 6	*W*	*2.408*	*-3.492*	*-2.761*	*-2.078*	*-1.553*	*-0.796*	*-0.312*	*0.495*	*1.636*
	M	2.384	-4.251	-3.773	-3.021	-2.475	-1.749	-1.17	-0.245	0.844
Item 7	*W*	*2.894*	*-5.171*	*-3.842*	*-1.992*	*-0.895*	*0.007*	*0.855*	*1.773*	*4.324*
	M	2.478	-2.773	-1.296	0.5	1.375	2.123	2.715	3.312	5.624
Item 8	*W*	*2.987*	*-6.914*	*-5.16*	*-3.465*	*-2.249*	*-1.31*	*-0.315*	*0.486*	*3.273*
	M	2.758	-4.819	-3.517	-1.732	-0.506	0.323	1.124	1.844	4.649
Item 9	*W*	*2.779*	*-7.505*	*-5.592*	*-3.751*	*-2.617*	*-1.829*	*-0.871*	*-0.148*	*2.07*
	M	3.375	-5.195	-3.681	-1.819	-0.572	0.44	1.464	2.239	5.014

Note. Item discrimination is denoted as "a". b_1 to b_8 signifies item locations and reflects the threshold level of the latent trait necessary to have at least a 50% probability of endorsing the next response choice. Item parameters reflect those obtained with the GRM model. Item parameters for women (*W*) appear in italics.

DIF Results.

Overall, MIMIC models identified Items 1, 4, and 8 of the SOI-R as having uniform DIF only. Items 5, 7, and 9 were identified as having both uniform and non-uniform DIF, indicating that the magnitude of advantage received by one group varied across the continuum. However,

non-uniform DIF is essentially an interaction, as with other statistical procedures, significant interactions are interpreted first and uniform DIF only interpreted in the absence of meaningful non-uniform DIF (de Ayala, 2009). Therefore, these items were considered to exhibit non-uniform DIF, along with Item 6, which exhibited soley non-uniform DIF. Only Items 2 and 3 were identified as being DIF free. MIMIC results, including standardized estimates and standard error for each direct effect are reported in Table 9 (uniform DIF) and Table 10 (non-uniform DIF). Item parameters and DIF results for are detailed for each item individually next.

Table 9. Uniform DIF for SOI-R Items.

Item	*β*	*SE*	*p*
1. With how many different partners have you had sex with in the past 12 months?	0.32	0.06	**.024**
2. With how many different partners have you had sexual intercourse on one and only one occasion?	-0.08	0.05	.121
3. With how many different partners have you had sexual intercourse without having an interest in a long-term committed relationship with this person?	0.09	0.07	.199
4. Sex without love is OK.	0.61	0.07	**< 001**
5. I can imagine myself being comfortable and enjoying "casual" sex with different partners.	-0.55	0.12	**< 001**
6. I do not want to have sex with a person until I am sure that we will have a long-term, serious relationship. (R)	-0.11	0.12	.388
7. How often do you have fantasies about having sex with someone you are not in a committed romantic relationship with?	-0.34	0.10	**< 001**
8. How often do you experience sexual arousal when you are in contact with someone you are not in a committed romantic relationship with?	0.58	0.09	**< 001**
9. In everyday life, how often do you have spontaneous fantasies about having sex with someone you have just met?	-0.18	0.09	**.034**

Note. Because men were coded as “0” and women as “1,”significant positive signed estimates indicate DIF favoring women, while negative signage indicates DIF favoring men. Significant uniform DIF is boldfaced.

Table 10. Non-Uniform DIF for SOI-R Items.

Item	*β*	*SE*	*p*
1. With how many different partners have you had sex with in the past 12 months?	0.11	0.13	.363
2. With how many different partners have you had sexual intercourse on one and only one occasion?	-0.15	0.08	.051
3. With how many different partners have you had sexual intercourse without having an interest in a long-term committed relationship with this person?	0.01	0.06	.820
4. Sex without love is OK.	-0.02	0.05	.736
5. I can imagine myself being comfortable and enjoying "casual" sex with different partners.	0.11	0.03	**.001**
6. I do not want to have sex with a person until I am sure that we will have a long-term, serious relationship. (R)	0.13	0.05	**.015**
7. How often do you have fantasies about having sex with someone you are not in a committed romantic relationship with?	0.21	0.04	**< .001**
8. How often do you experience sexual arousal when you are in contact with someone you are not in a committed romantic relationship with?	-0.02	0.03	.609
9. In everyday life, how often do you have spontaneous fantasies about having sex with someone you have just met?	-0.20	0.03	**< .001**

Note. Because men were coded as "0" and women as "1,"significant positive signed estimates indicate DIF favoring women, while negative signage indicates DIF favoring men. Significant non-uniform DIF is boldfaced.

Factor 1: Behavior

Item 1. The results of the MIMIC model testing Item 1, "*With how many different partners have you had sex with in the 12 months?*" for gendered DIF revealed statistically significant uniform DIF was present in the item (p = .024) after controlling for mean differences on the latent trait. Given that women were coded as "*1*", the positive estimate for the regression relating Item 1 to Gender indicates that the DIF favored women, with the amount of θ needed for a 50% or greater probability of endorsing the next higher response choice being lower for men

uniformly across the continuum. Thresholds ranged from -5.36 to 2.07 for women and -4.64 to 1.65 for men (all thresholds reported in Table 8). For men, peaks were higher for each response category when examining the data for women. Item plots for Item 1 revealed that women with *θs* at the low end of the spectrum (*θ* of roughly -5 and lower) were most likely to choose the 1st response option, indicating having had no sexual partners in the past 12 months. Conversely, those at the opposite end (*θ* of 2 or higher) were most likely to pick response option 9, indicating having had 10 or more partners. Response category 2, indicating one sexual partner, was another likely choice for respondents with lower than average *θ* , and was in fact the most chosen option for both women (*n* = 453, 44.3%) and men (*n* = 424, 44.8%). As shown in Table 8, women had a higher probability than men of endorsing a response category across the board, after being matched on sociosexuality. Item parameters for Item 1 indicated that the item was a slightly better discriminator for women (1.24) than for men (1.04). Item information plots for both men and women (see Appendix), revealed that Item 1 provided the most information for those with higher than average *θ*, spanning from *θ* of about -6 to 7, with maximum information provided for *θ* of about 4. The amount of information provided was higher for women than it was for men, as the item discrimination values also suggest.

In sum, Item 1 exhibited uniform DIF indicating that the probability that women would choose a higher response choice on the item was higher than it was for men, despite men and women being matched on sociosexuality. The item was revealed to provide the most information for women with above average levels of the latent trait.

Results for Item 1 further indicated no statistically significant relationship between respondents' gender and number of sexual partners during the past 12 months. However, significant relationships were revealed between respondents' race and Kinsey scores, such that

identifying as white and non-heterosexual predicted having higher levels of the trait (i.e., unrestricted orientations) than non-whites and heterosexual individuals. Means and standard deviations for all items are presented according to gender in Table 3.

Item 2. The results of the MIMIC model revealed a statistically non-significant result for Item 2, " *With how many different partners have you had sexual intercourse on one and only one occasion?"* for both uniform (p = .121) and non-uniform DIF (p = .051), indicating that reporting the number of partners was not influenced by Gender. Thresholds for men and women ranged from roughly -8 to 1.5 and discrimination values suggested that the item was an excellent discriminator for both men (4.01) and women (4.43). Item 2 provided information for a smaller range of θ than Item 1 (from θ of approximately -2 to 3 for both men and women); however, the amount of information was larger, with maximum information provided for individuals with θ slightly above average at 1. In sum, Item 2 did not exhibit DIF and was found to provide the most information about men and women with slightly higher than average levels of the latent trait.

In terms of impact, results of the model did not reveal a significant relationship between Gender and the latent trait (β = - 0.02, p = .583). However, significant relationships between the latent trait, Race (β = 0.23, p < .001) and Kinsey scale scores (β = 0.08, p < .001) were revealed, such that white, non-heterosexual respondents had orientations that were more unrestricted.

Item 3. Item 3, "*With how many different partners have you had sexual intercourse without having an interest in a long-term committed relationship with this person?"* was also found to be free of uniform DIF (β = 0.09, p = .199) and non-uniform DIF (β = 0.01, p = .820), indicating that respondents' gender did not influence responses. Item thresholds for men and women ranged from approximately –7 to 1.8, suggesting that lower than average levels of the

latent trait were necessary for most item response choices. Plots for Item 3 revealed that category 2 was most probable for men and women with lower than average θ, and categories 6 and 8 were most probable for those with θ slightly higher than average. Item 3 was also shown to be an excellent discriminator for both men (4.17) and women (4.34) and provided information for men and women with θ between -1 and 3. Item 3 provided the most information about respondents out of all nine items, with the maximum amount of information provided for respondents with θ of 1. In sum, Item 2 did not exhibit DIF and was found to provide a lot of information about men and women with slightly higher than average levels of the latent trait.

Results of the MIMIC model relating θ to Gender and the covariates further revealed a non-significant relationship between sociosexuality and Gender (β = -0.60, p = .214), indicating there were no significant mean differences between men and women in the number of sexual partners each had with no intention in forming a long-term relationship. However, Race (β = 0.23, $p < .001$) and Kinsey scale scores (β = 0.08, $p < .001$) indicated that respondents who were non-heterosexual and white were more likely to have unrestricted orientations than non-white heterosexuals.

Factor 2: Attitude

Item 4. The results for Item 4 revealed a statistically significant result for uniform DIF (β = .61, $p < .01$) favoring women, indicating they had a higher probability than men of expressing agreement with the statement "*Sex without love is ok*" after being matched on sociosexuality. Results for non-uniform DIF (β = -0.02, p = .736) were statistically nonsignificant. Thresholds ranged from -0.54 to 4.24 for women, with only response category 1 being the most likely choice for below average θ, suggesting that, compared to the other items of the SOI-R, Item 4 required higher levels of the latent trait (unrestricted sociosexual orientations) to endorse. This response pattern extended to men, but with thresholds ranging from -0.41 to 5.14, indicating that they

needed higher levels of the latent trait for the probability of endorsing a given response to reach 50% or greater. Discrimination parameters suggested that Item 4 was good at differentiating between restricted and unrestricted orientations in both men (3.06) and women (2.41). For men, the item provided the most information for those with θ ranging from -4 to 2, with the maximum amount of information provided for men with average sociosexuality (θ of 0). For women, Item 4 provided information for those with θ from -3 to 2. Maximum information provided by Item 4 was higher for women than it was for men, however, the item also provided the most information for θ of 0. In sum, Item 4 exhibited both uniform and non-uniform DIF, with non-uniform DIF favoring women. The item provided the most information for women with typical levels of the trait.

Results assessing impact for Item 4 revealed statistically significant relationships between the latent trait and Gender (β = -0.74, p < .001), Race (β = 0.68, p < .001), and Kinsey scale scores (β = 0.23, p < .001) such that being male, white, and non-heterosexual was associated with higher agreement with the item.

Item 5. Results for the MIMIC model assessing whether DIF was present in the item "*I can imagine myself being comfortable and enjoying 'casual' sex with different partners*" revealed statistically significant uniform DIF (β = -0.55, p < .001) and non-uniform DIF (β = 0.11, p = .001). As previously noted, the significant interaction takes precedence and therefore, uniform DIF results were not interpreted. Thresholds for Item 5 indicated that even when matched on sociosexuality scores, it was easier for women, whose thresholds ranged from -3 to 3.68, to endorse responses across the continuum than it was for men, whose thresholds ranged from -1.17 to 5.19. Closer examination of thresholds (Table 8) suggest that the discrepancy between men and women's response choices was smaller when assessing respondents at below

average θ and greater when assessing the trait at the higher end of the continuum (i.e., the 5th threshold and above). Discrimination parameters revealed that Item 5 was a good discriminator for men (3.67) and even better for women (4.02). Item 5 was found to provide the most information for men with θ between -3 and 2, with maximum information again provided for those with average θ at 0. For women, Item 5 provided the most information for those with θ between -2 to 2, maxing out at about 1. As indicated by discrimination values, Item 5 provided more information, overall, for women.

In sum, Item 5 exhibited both uniform and nonuniform DIF favoring women, indicting they needed lower levels of the latent trait than men to endorse higher item responses. Item 5 provided a lot of information, overall, but provided the most for women with slightly higher than average levels of the latent trait.

Assessment of the item for impact further revealed significant relationships between the latent trait and Gender (β = -0.42, p < .001), Race (β = 0.66, p < .001), and Kinsey scores (β = 0.22, p < .001), such that white, non-heterosexual men were more likely to have expressed higher levels of agreement with the item.

Item 6. Results for Item 6 indicated the presence of non-uniform DIF(β = 0.13, p = .015), while being free of uniform DIF (β = -0.11, p = .388). Thresholds for women ranged from -3.5 to 1.6, while thresholds for men ranged from -4.23 to 0.84, indicating that higher levels of the latent trait were not necessary to endorse the reverse coded statement "*I do not want to have sex with a person until I am sure that we will have a long-term, serious relationship*." Closer examination of the thresholds (Table 8) revealed that there was a higher probability of women endorsing each response category than men, across the continuum. For men and women, response categories 1 and 9 were again the most probable choices for those at each end of the

sociosexuality continuum. The 7th response category was the most likely for men with average θ, while the 8th category appeared to be the most likely for women. Item 6 exhibited good discrimination for both men (2.38) and women (2.41). For men and women, Item 6 provided the most information for θ between -3 and 3, with maximum info provided for women with θ of 1. In sum, Item 6 exhibited uniform DIF favoring women and provided the most information about respondents with slightly higher than average levels of the trait.

Gender (β = -0.62, $p < .001$), Race ($\beta = 0.63$, $p < .001$) and Kinsey scores ($\beta = 0.21$, $p < .001$) were all significantly related to the latent trait, with identifying as white, male, and non-heterosexual predicting unrestricted orientations.

Factor 3: Desire

Item 7. Results for Item 7,*"How often do you have fantasies about having sex with someone you are not in a committed romantic relationship with?"* revealed the presence of both non-uniform DIF($\beta = 0.21$, $p < .001$) and uniform DIF ($\beta = -0.34$, $p < .001$), with only non-uniform DIF being reported further. Thresholds for women ranged from -5.17 to 4.32 and from -2.78 to 5.62 for men and discrimination values were revealed to be good for both men (2.48) and women (2.89). Item 7 was revealed to provide the most information for men with θ ranging from -4 to 3 and maxing out at about 0. For women, the most information was provided for θ between -3 and 4, with the most provided for θ of 2 and overall, was slightly higher than for women. In sum, Item 7 exhibited both uniform and non-uniform DIF that favored women and provided the most information for women with higher than average levels of the trait.

Gender ($\beta = -1.14$, $p < .001$), Race ($\beta = 0.376$, $p < .001$) and Kinsey scores ($\beta = 0.159$, $p < .001$) were again revealed to be significantly related to the latent trait, with white, non-heterosexual men endorsing response choices indicative of higher levels of agreement with the item.

Item 8. The results for Item 8, "*How often do you experience sexual arousal when you are in contact with someone you are not in a committed romantic relationship with?"* revealed statistically significant uniform DIF (β =.58, $p < .01$) favoring women and a non-significant result for non-uniform DIF (β = -0.02, p = .609). Thresholds ranged from -6.91 to 3.27 for women and -4.81 to 4.67 for men, suggesting that, after being matched on the latent trait, women had a higher probability of endorsing all item response categories across all levels of θ. All response options except for the 3rd and 5th were most probable for a segment of θ for women. For men, this pattern was also observed, except the 4th response option was included as one that was never the most likely at any level of the latent trait. Item 8 had good discrimination values for both men (2.76) and women (2.99) and provided the most information for men whose sociosexuality ranged from θ of -4 to 4, with the maximum amount of information provided at θ of about 2. For women, the most information was provided for those with θ of about -3 to 4, and also maxed out at about 2. As with Item 7, Item 8 also provided slightly higher information for respondents who were women, rather than men. In sum, Item 8 exhibited uniform DIF that favored women, indicating they required lower levels of the trait than did men to endorse higher response categories. The item provided the most information for women with higher than average levels of the latent trait.

Results assessing impact revealed statistically significant relationships between the latent trait and Gender (β = -1.53, $p < .001$), Race (β = 0.44, $p < .001$), and Kinsey scores (β = 0.18, $p < .001$) such that identifying as a white, non-heterosexual male predicted higher agreement with the item.

Item 9. Lastly, results for Item 9, "*In everyday life, how often do you have spontaneous fantasies about having sex with someone you have just met?"* revealed the presence of both

non-uniform DIF(β = -0.20, p < .001) and uniform DIF (β = -0.18, p = .034) favoring men, with only non-uniform DIF receiving interpretation. Thresholds for women ranged from -7.5 to 2.07 for women and from -5.20 to 5.01 for men. All but the 4th option were most probable for a segment of θ for men. For women, all response categories but the 3rd and 5th were most probable for a segment of θ. Item 9 was also revealed to have high discrimination values for women (2.78) and were even higher for men (3.38). For men, Item 9 provided the most information for those with θ between -3 and 3, maxing for those with θ of about 1. For women, Item 9 provided slightly less information, overall for θ ranging between -2 and 5, maxing out at around 2. In sum, Item 9 exhibited both uniform and non-uniform DIF, which favored men. The item provided the most information about men whose sociosexuality was slightly higher than average.

Gender (β = -1.32, p < .001), Race (β = 0.43, p < .001) and Kinsey scale scores (β = 0.18, p < .001) were again revealed to be significantly related to the latent trait, with white, non-heterosexual men endorsing higher agreement with the item.

Summary of Results

In sum, results revealed no mean differences between men and women on Factor 1: Behavior. There were slight mean differences found for Factor 2: Attitude, with men's average being slightly higher than women's, and larger mean differences on Factor 3: Desire—again with men's average revealed as being higher. A CFA revealed the three-factor solution was adequate, with most factor loadings exceeding .80. Item 1 had the lowest factor loading at .49. Results of MIRT analyses for men and women's responses indicated that the GRM fit the data best. Item parameters were obtained and revealed that the items of the SOI-R all had discrimination values that were higher than 1, with most exceeding 2, indicating that the items were good at differentiating between individuals with varying sociosexual orientations. Items 2 and 3 had the highest discrimination values and provided the most information about respondents, whereas

Item 1 was found to provide the least amount of information. These results further revealed that the items provided information for an evenly distributed range of θ, with the maximum amount of information being provided for average θ (or close to average) for the majority of items. The results of the MIMIC models identified uniform DIF on Items 1, 4, and 8. Items 5, 6, 7, and 9 were identified as having non-uniform DIF, while Items 2 and 3 were found to be DIF free. Item 9 exhibited DIF favoring men, whereas the remainder of items where DIF was present favored women. These results suggest that the hypothesis that DIF on the behavior subscale would exhibit DIF favoring men was not supported. Additionally, the prediction that the items comprising the attitude subscale would exhibit DIF favoring men was also not supported with the MIMIC model results.

CHAPTER FIVE: DISCUSSION

The purpose of this study was to assess the items of the SOI-R, an instrument used to assess sociosexual orientation, for gendered DIF. Because the SOI-R is an instrument commonly used in the social sciences, determining that the items were performing equally across groups has important implications for the validity of research claims regarding sociosexuality, particularly those involving gender differences.

Based on the findings from previous research indicating that social pressures result in inaccurate responding to questions pertaining to sex (e.g., Alexander & Fisher, 2003; Fenton et al.,2001; Krumpal, 2013), it was expected that items measuring sociosexual behavior and attitudes would be found to be performing differently for men and women. More specifically, due to consistent findings throughout the literature stating that men tend to *overreport* number of sexual partners whereas women have been found to *underreport* theirs, it was expected that the three items assessing the number of past sexual partners would exhibit DIF favoring men (Alexander &Fisher, 2003; Fenton et al., 2001; Mitchell et al., 2018). Additionally, because of

social pressures for women to appear as chaste and only express interest in sex within the context of a committed, long-term monogamous relationship, it was predicted that three items of the sociosexual attitudes subscale would also exhibit DIF favoring men (Emmerink et al., 2016). There were no specific predictions regarding items 7 through 9, which comprised the sociosexual desire subscale, however these items were examined for DIF and explored as a research question.

Descriptives results for the SOI-R revealed that, consistent with previous research (e.g., Penke & Asendorpf, 2008), mean gender differences on items were greatest for those comprising the Desire subscale, followed by those of the Attitude subscale, and least for the Behavior subscale. Theoretically, this observed pattern is a result of both biological and cultural factors. Innate differences between men and women as a result of evolutionary pressures shape the larger discrepancies seen in sexual desires, whereas social and cultural factors influence attitudes towards sex. Because men and women within a given population are typically products of the same environment, these differences are smaller, but exist nonetheless as a result of social scripts such as the SDS (Penke & Asendorpf, 2008; Schmitt, 2005; Simpson & Gangestad, 1991; Zaikman & Marks, 2017). Although previous research asserts that sociosexual behavior in heterosexual men and women should align in the number of sexual partners as a result of simple mathematical realities, as previously noted, men and women's self-reports of past sexual partners are often misaligned for a variety of reasons, including a desire to conform to gendered norms (Mitchell et al., 2018). As the results of this study revealed, however, men and women did not differ in the self-reporting of past sexual partners. This could be due to a variety of reasons, including greater honesty in responding resulting from the fact that the study was administered online and responses were anonymous. That respondents indicated past sexual partners using a 9-point scale may have also contributed to this absence. Had respondents been required to provide

an open-ended estimate, differences may have been seen. It is also worth noting that mean differences on item 2 and 3, which assessed the number of different partners respondents had sexual intercourse with in the context of an uncommitted relationship, did reveal slightly higher averages for men prior to the removal of data for people who identified as asexual, which were primarily women (17 versus 6 men). The lack of discrepancy between men and women's reported number of sexual partners could also have been due to selection bias, with people who chose to participate in a study examining attitudes towards sexual experiences and behaviors feeling less pressure to respond in ways that exemplified traditional norms—a possibility that is expanded upon in greater detail in a following paragraph.

Results of the MIRT analyses revealed that, overall, the items provided information on men and women's sociosexuality for a well-distributed range of θ. Many of the items provided the maximum amount of information for respondents with average, or close to average, θ, meaning that their sociosexual orientation fell somewhere in the middle of restricted versus unrestricted. Item 1 was shown to have provided the least amount of information on sociosexuality about respondents and had the lowest factor loading of all nine items. This is not surprising given that the remaining eight items were more specific to behaviors, attitudes, and desires in the context of casual sex. Additionally, most men and women reported currently being in a relationship, which could account for the lower average number of partners during the past year, regardless of total SOI-R scores. Conversely, Items 2, 3 and 5, which assessed the number of past casual sexual partners and anticipated comfortability with the idea of enjoying casual sex with new and different partners, provided a relatively high amount of information for both men and women, with Item 3 providing the most.

The results of the MIMIC model assessing DIF revealed that, overall, the hypothesis that items comprising the Behavior subscale would exhibit DIF favoring men was unsupported. Had this been the case, results would have indicated that despite having identical scores on the SOI-R, men were more likely than women to have chosen a response indicative of having had a greater number of past sexual partners. Only Item 1 was found to demonstrate DIF; however, the sign of the estimate regressing the item on Gender indicated that the DIF favored women and not men, as had been predicted. This suggests that when matched on the latent trait, women had the higher probability of endorsing a response choice indicating a greater number of sexual partners during the past year. Similarly, the items comprising the attitude subscale did not support the hypothesis that the items would exhibit DIF favoring men. These items also exhibited DIF favoring women; however, it is unclear why this is the case. One possible explanation is that because women are cognizant of the fact that they are expected to hold negative attitudes towards and refrain from casual sex, female respondents were eager to dispel that stereotype and felt compelled to provide responses indicating greater favorability towards casual sex. For example, it is plausible that many women who opted to take the survey considered themselves to hold positive views about sex. Consequently, when responding to Item 4, "*Sex without love is ok*," which demonstrated the largest effect according to the standardized estimate, women who perceive themselves as sex-positive may have felt especially obliged relative to men to indicate agreement with the item. However, it is also possible that men who partook in the study were motivated to present themselves as antithetical to the stereotype of being overly pursuant of casual sex.

The possibility that women experienced internal pressure to indicate higher levels of agreement with items is particularly compelling when considering selection bias for sex studies.

As previously described in the methodology section, the questionnaire in which the SOI-R was embedded contained numerous other items assessing attitudes towards and experiences with group sex. Given that several items required respondents to indicate their moral attitudes, anticipated pleasure, and interest in a variety of threesome scenarios, respondents who made it to the portion of the survey where the items of the SOI-R were positioned (about 100 items in) were likely considerably more comfortable with and held more positive attitudes towards casual sex than people who did not take the survey, or that quit before reaching the items of the SOI-R. Indeed, previous research has found that volunteers for sexuality studies do hold less traditional attitudes towards sex, have higher sexual self-esteem, and may be more sexually experienced (Lehmiller, 2018;Weiderman, 1999). So, although there were people who reported negative view towards non-traditional forms of sex (e.g., about 20% chose responses indicating negative views towards both men and women who participated in threesomes), it is likely that these views occurred less frequently among respondents than would typically be seen throughout the population (Lehmiller, Kirkeby, & Cipriano, 2018). Consequently, the results of this study cannot be said to be representative of the general population, as individuals that are most uncomfortable with sex, and perhaps hold negative views, are less likely to participate in sex studies and are therefore not represented. Rather, these results provide insight into the attitudes and behaviors of men and women who are willing to participate in a study about sex—and more non-traditional forms of sex, at that.

Items 7 through 9, which were explored as research questions, were also shown to exhibit DIF, with items 7 and 8 favoring women and Item 9 favoring men. Again, a possible explanation for the finding is that women were motivated to respond in a way consistent with the self-image as being sex-positive. Because Item 9 was the only item to exhibit DIF favoring men, it is difficult

to ascertain the accuracy of this finding, but a possible explanation is that men and/or women *were* motivated to conform to traditional gender norms stating men should convey strong sexual desire, while women should convey the opposite (Emmerink et al., 2016).

Overall these findings suggest that observed gender differences on the SOI-R may actually be *larger* than initially thought, rather than smaller as hypothesized. Additionally, women who have been scored as having unrestricted sociosexual orientations may exhibit traits or behaviors more consistent with orientations more on the restricted side. As noted, however, causality cannot be inferred from a DIF analysis, suggesting the findings that many items exhibited DIF favoring women is an ideal area for future research to explore.

Lastly, MIMIC results assessing the influence of demographic covariates on sociosexuality revealed that race and Kinsey scale scores were significantly related to the latent trait for all items, indicating that identifying as white and non-heterosexual predicted having an unrestricted sociosexual orientation. Given that this relationship was frequently observed, it is recommended that future research also assess the items of the SOI-R for DIF based on race and sexual orientation.

Strengths and Limitations

There are notable limitations to this study. One limitation is that item purification was not performed on the instrument when testing for DIF, increasing the chance that a Type I error occurred. While performing item purification by removing items flagged for DIF from the scale prior to retesting the remaining items would have resulted in a pure set of anchor item, there were too few of items in each subscale to do so. Additionally, having only three indicator items for each factor may have compounded this issue. Consequently, there is a risk that some of the items that were found to exhibit DIF were actually DIF free. Although the MIMIC model is quite flexible in the conditions that it can assess DIF, it does not offer a means to assess effect sizes for

DIF in ordinally scored items. As a result, the only means of assessing the magnitude of DIF was the *p* values and standardized estimates, which may not be as reliable and run the risk of falsely flagging items for DIF. Another limitation was that although some demographic variables were included in the MIMIC models as covariates, others such as age, relationship status, and education level were not. Therefore, it is possible that these additional variables were significantly related to the latent trait and should have also been controlled for in the models. It is recommended that future research examine the SOI-R items for DIF based on additional demographic variables.

Finally, as previously detailed, it is likely that responses reflected attitudes and behaviors of men and women with greater comfortability with sex than the general population. Although the means for item subscales were comparable to those obtained by other researchers (e.g., Penke & Asendorpf, 2008), such comparisons do not eliminate the problem of selection bias. Undoubtedly, the presence of this bias is not ideal—particularly when attempting to assess group differences in item performance; however, it is a limitation that is pervasive throughout sex research, with no clear-cut solution.

Despite these limitations, however, this research contributes to the literature by being the first known study to conduct a full IRT analysis on the items of the SOI-R, therefore providing an assessment of how each of the items perform. Additionally, this is the first known study to assess the items for DIF, therefore providing insight into how the items are performing for members of different groups.

Practical Implications for Research

Overall, the results of this study suggest that the items of the SOI-R do provide information on respondent's sociosexuality and that they do so for all levels of the trait (i.e., those with restricted, average, and unrestricted orientations). However, the majority of items

were found to provide the most information for individuals with typical, or close to typical, levels of the trait. Therefore, it is unknown how the scale would perform in populations with especially restricted or unrestricted sociosexual orientations. Items 2, 3, and 5 provided the most information on sociosexuality and should therefore be given primary consideration in situations where researchers aim to assess willingness to engage in casual sex with a minimal number of items. Items 2 and 3 were also found to be DIF free, further suggesting they should be given first consideration in relevant research. Conversely, Item 1 did not provide a lot of information on sociosexuality, suggesting that including it on questionnaires assessing sociosexuality may not be as important.

Although the items were found to perform well at assessing sociosexuality for both men and women, they did perform differently for each gender. Contrary to predictions, women had a higher probability of endorsing response categories compared to men on most items, with six out of the seven items exhibiting DIF providing an advantage to women. But, as noted, this study was not able to discern why these differences occurred or specific effect sizes for DIF on items where it was present. Because the expectation for measurement scales is that they provide an accurate and unbiased assessment of the construct of interest, these results suggest that use of the SOI-R to assess sociosexuality could be problematic in some instances. More specifically, researchers should exercise caution when drawing conclusions about what the scale tells us about gender differences in sociosexuality. That said, the SOI-R is most often employed in tandem with other personality and attitudinal assessments to examine how sociosexual orientation relates to various personality traits or predicts attitudes and behaviors. Therefore, some consideration should be given when drawing conclusions on how women's sociosexuality is associated with attitudes and behaviors—even when gender differences are not relevant. For example, because

results indicated that women needed lower levels of the latent trait (i.e., orientations that were more restricted) to endorse item responses than did men with equivalent SOI-R scores, the items and scales may not accurately predict behaviors (e.g., having a greater number sexual partners or sexual infidelity) to the extent that they otherwise would. Put another way, women who are scored as having orientations that are more unrestricted may not hold as liberal views towards, desire, or engage in casual sex to the extent that might be expected given their SOI-R scores.

This is not to suggest that the SOI-R is completely invalid and should not be used in research, however—especially since most items were found to provide a good deal of information on the trait the scale was designed to assess. Rather, until more research sheds further light into why the items are performing differently, results should be interpreted with some degree of caution, particularly regarding women's sociosexuality. Furthermore, the presence of DIF does not necessarily mean that the items of the SOI-R aren't assessing sociosexuality in the way that they were intended. Although the specific hypotheses for this study were not supported and the results were, in fact, in the other direction, it has been suggested throughout this study that social factors may be shaping response patterns, not the instrument itself.

Researchers who use the instrument should account for the gender discrepancies in item performance before making claims regarding the trait, however, especially when these claims involve gender differences in sociosexuality. To achieve this, there are several approaches that researchers might employ (Teresi, Ramirez, Jones, Choi, & Crane, 2012). First, future research should determine a precise estimate of the magnitude of DIF for each item. Because the sample size for this study was fairly large, which can lead to a statistically significant result where there is, in fact no DIF, researchers may want to first replicate this study using smaller and diverse

samples and with alternative approaches where an effect size for DIF can be obtained (e.g., logistic regression). It is possible that some of the contaminated items may be found to exhibit inconsequential DIF, or no DIF altogether. For items where large DIF occurs, researchers may want to remove these items before administering the scale, particularly if the goal is to compare results between men and women. Alternatively, researchers may attempt to correct for DIF by adjusting the means accordingly to compensate for women's advantage on items. A final option would be for experts to determine why the items on the SOI-R provide an advantage to women, however, given that the DIF may be occurring as a result of social factors that are potentially outside the control of psychologists or other researchers, this option may not prove beneficial. Regardless of the path that future research takes on this issue, the end goal is to obtain accurate and informative insight into human traits. The results of this study have uncovered a barrier to achieving this end goal, but in doing so, has put us one step closer.

References

Ackerman, T. A. (2005). Multidimensional item response theory modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), Contemporary psychometrics (pp. 3-26). Mahwah, NJ: Lawrence Erlbaum.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov, & F. Csaki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Alexander, M. G., & Fisher, T. D. (2003). Truth and consequences: Using the bogus pipeline to examine sex differences in self-reported sexuality. *Journal of Sex Research, 40*(1), 27-35.

Alicke, M., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. European Review of Social Psychology 20, 1-48.

Allison, R., & Risman, B. J. (2013). A double standard for "hooking up": How far have we come toward gender equality? *Social Science Research, 42*(5), 1191-1206.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice Hall.

Baumeister, R. F., Catanese, K. R., & Vohs, K. D. (2001). Is there a gender difference in strength of sex drive? Theoretical views, conceptual distinctions, and a review of relevant evidence. *Personality and social psychology review*, *5*(3), 242-273.000

Baumeister, R. F., & Finkel, E. J. (Eds.). (2010). Advanced social psychology: The state of the science. New York, NY, US: Oxford University Press.

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test,

and logistic regression. *Frontiers in Education*. Retrieved from https://www.frontiersin.org/articles/10.3389/feduc.2017.00051/full.

Buss, D. M. (1998). Sexual strategies theory: Historical origins and current status. *Journal of Sex Research*, *35*(1), 19-31.

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: an evolutionary perspective on human mating. *Psychological review*, *100*(2), 204-232.

Chalmers, R., P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. doi: 10.18637/jss.v048.i06

Chun, S., (2014). *Using MIMIC Methods to Detect and Identify Sources of DIF among Multiple Groups*. (master's thesis). Retrieved from https://scholarcommons.usf.edu/etd/5352.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: issues and practice, 17*(1), 31-44.

de Ayala, R. J. (2009). *Methodology in the social sciences. The theory and practice of item response theory.*New York, NY, US: Guilford Press.

De Leo, J. A., Van Dam, N. T., Hobkirk, A. L., & Earleywine, M. (2011). Examining bias in the impulsive sensation seeking (ImpSS) Scale using Differential Item Functioning (DIF)–An item response analysis. *Personality and Individual Differences, 50*(5), 570-576.

Dodou, D., & de Winter, J. C. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, *36*, 487-495.

Emmerink, P. M., Vanwesenbeeck, I., van den Eijnden, R. J., & ter Bogt, T. F. (2016). Psychosexual correlates of sexual double standard endorsement in adolescent sexuality. *The Journal of Sex Research*, *53*(3), 286-297.

Fenton, K. A., Johnson, A. M., McManus, S., & Erens, B. (2001). Measuring sexual behaviour: methodological challenges in survey research. *Sexually Transmitted Infections, 77*(2), 84-92.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295.

Finch, W. H, & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*(4), 565-582.

Finch, W.H. & French, B.F. (2015). *Latent Variable Modeling with R*. New York: Routledge.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Hyde, J. S. (2005). The gender similarities hypothesis. *American psychologist, 60*(6), 581-592.

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65*, 373-398.

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the mini-mental state examination: effects of differential item functioning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 57*(6), P548-P558.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351a), 631-639.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*(1), 51-66.

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). *Sexual behavior in the human male.* Philadelphia: Saunders.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*(4), 2025-2047.

Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement*, *77*(4), 545-569.

Lehmiller, J., J. (2018). *Tell Me What You Want.* New York City, NY: De Capo Press

Lehmiller, J. J., Kirkeby, K., & Cipriano, A. (2018). [Interest in Threesomes]. Unpublished raw data.

Lord, F. (1952). A Theory of Test Scores (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23,* 157–162. https://doi.org/10.1111/j.1745- 3984.1986.tb00241.x

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement, 27*(5), 372-379.

Marks, M. J., & Fraley, R. C. (2005). The sexual double standard: Fact or fiction? *Sex Roles, 52*(3-4), 175-186.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Mazor, K. M., Hambleton, R. K., and Clauser, B. E. (1998). Multidimensional DIF analyses: the effects of matching on unidimensional subtest scores. *Appl. Psychol. Measure.* 22, 357–367. doi:10.1177/014662169802200404

Mitchell, K. R., Mercer, C. H., Prah, P., Clifton, S., Tanton, C., Wellings, K., & Copas, A. (2018). Why Do Men Report More Opposite-Sex Sexual Partners Than Women? Analysis of the Gender Discrepancy in a British National Probability Survey. *The Journal of Sex Research, 00*(00), 1‑8.

Mitchelson, J. K., Wicher, E. W., LeBreton, J. M., & Craig, S. B. (2009). Gender and ethnicity differences on the Abridged Big Five Circumplex (AB5C) of personality traits: A differential item functioning analysis. *Educational and Psychological Measurement, 69*(4), 613-635.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(*2*),* 159–176.

Muthen, B. O., & Satorra, A. (1995). Technical aspects of Muthen's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60*(4), 489-503.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4), 315-328.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Sage Publications.

Penke, L., & Asendorpf, J. B. (2008). Beyond global sociosexual orientations: a more differentiated look at sociosexuality and its effects on courtship and romantic relationships. *Journal of Personality and Social Psychology, 95*(5), 1113-1135. doi:10.1037/0022-3514.95.5.1113

Petersen, J. L., & Hyde, J. S. (2010). A meta-analytic review of research on gender differences in sexuality, 1993–2007. *Psychological bulletin*, *136*(1), 21.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing* Vienna, Austria: R Foundation for Statistical Computing.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory. New York: Springer.

Schlenker, B. R. (1980). *Impression management*. Monterey, CA: Brooks/Cole Publishing Company.

Schmitt, D. P. (2005). Sociosexuality from Argentina to Zimbabwe: A 48-nation study of sex, culture, and strategies of human mating. *Behavioral and Brain Sciences, 28*, 247–275.

Schwarz, G. (1978). *Estimating the dimension of a model. Annals of Statistics, 6* , 461-464.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Simpson, J. A., & Gangestad, S. W. (1991). Individual differences in sociosexuality: Evidence for convergent and discriminant validity. *Journal of Personality and Social Psychology*, *60*(6), 870-833.

Simpson, J. A., Wilson, C. L., & Winterheld, H. A. (2004). Sociosexuality and Romantic Relationships. In J. H. Harvey, A. Wenzel, & S. Sprecher (Eds.), *The handbook of sexuality in close relationships* (pp. 87-112). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Sprecher, S., Treger, S., & Sakaluk, J. K. (2013). Premarital sexual standards and sociosexuality: Gender, ethnicity, and cohort differences. *Archives of Sexual Behavior*, *42*(8), 1395-1405.

Suh, Y., & Cho, S. J. (2014). Chi-square difference tests for detecting differential functioning in a multidimensional IRT model: A Monte Carlo study. *Applied Psychological Measurement*, *38*(5), 359-375.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44* (11), S152-S170.

Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S., & Crane, P. K. (2012). Modifying measures based on differential item functioning (DIF) impact analyses. *Journal of aging and health*, *24*(6), 1044-1076.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer

(Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Trivers, R. L. (1972). Parental Investment and Sexual Selection. In B. Campbell (Ed.), *Sexual Selection and the Descent of Man, 1871-1971* (pp. 136-179). Chicago, IL: Aldine.

Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*(5), 713-731.

Wiederman, M. W. (1999). Volunteer bias in sexuality research using college student participants. *Journal of Sex Research*, *36*(1), 59-66.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1-27.

Wood, W., & Eagly, A. H. (2002). A cross-cultural analysis of the behavior of women and men: Implications for the origins of sex differences. *Psychological Bulletin, 128*(5), 699-727. http://dx.doi.org/10.1037/0033-2909.128.5.699

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. In *Advances in experimental social psychology* (Vol. 46, pp. 55-123). Academic Press.

Zaikman, Y., & Marks, M. J. (2017). Promoting theory-based perspectives in sexual double standard research. *Sex Roles*, *76*(7-8), 407-420.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*.

APPENDIX

Demographics for Men

Education	*n*	*Percent*
Did not complete high school	7	.7
High School/GED	88	9.2
Some college, no degree	307	32.0
Associate's degree	81	8.4
Bachelor's degree	242	25.2
Master's degree	162	16.9
Doctoral or advanced professional	72	7.5
Race		
African American or Black	32	3.4
Asian or Pacific Islander	39	4.1
White or European American	785	82.2
Hispanic	51	5.3
Native American or Alaskan	7	.7
Biracial or Multiracial	19	2.0
Other	22	2.3

Demographics for Women

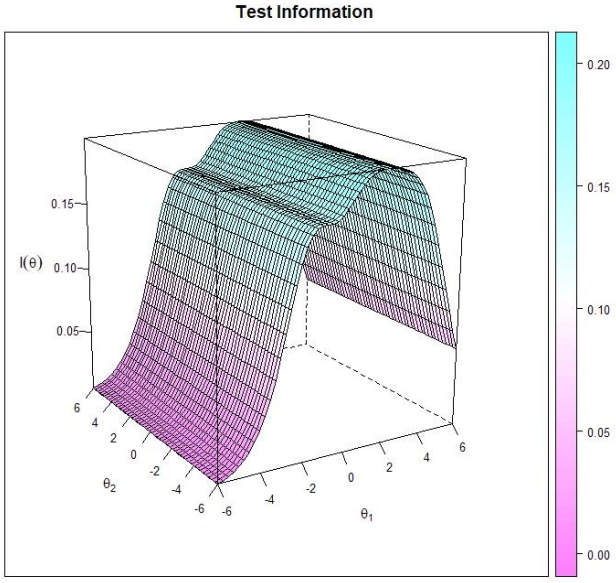
Education	*n*	*Percent*
Did not complete high school	7	.7
High School/GED	121	11.8
Some college, no degree	389	38.0
Associate's degree	113	11.0
Bachelor's degree	263	25.7
Master's degree	92	9.0
Doctoral or advanced professional	38	3.7
Race		
African American or Black	84	8.2
Asian or Pacific Islander	54	5.3
White or European American	707	69.2
Hispanic	89	8.7
Native American or Alaskan	6	.6
Biracial or Multiracial	48	4.7
Other	34	3.3

Item Response Theory Analysis of Items on the SOI-R (GRM) for Total Participants

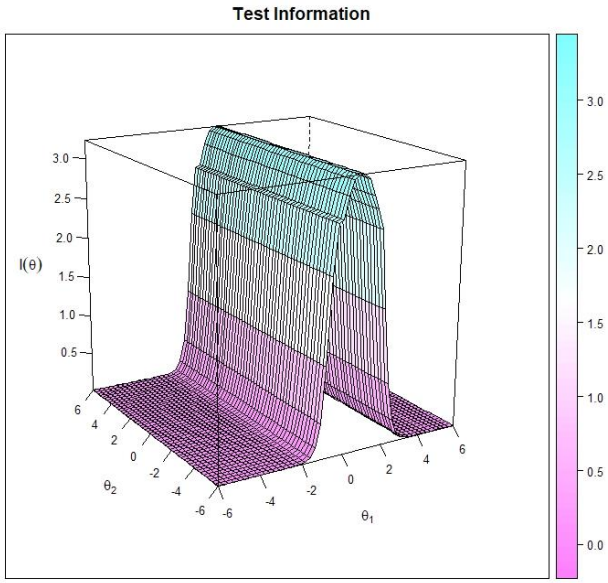
	a	b_8	b_7	b_6	b_5	b_4	b_3	b_2	b_1
Item 1	1.145	-4.964	-3.77	-2.284	-2.284	-1.828	-1.321	-0.68	1.835
Item 2	4.124	-8.243	-6.341	-5.152	-4.032	-3.263	-2.333	-0.96	1.415
Item 3	4.303	-7.432	-5.349	-4.329	-3.183	-2.247	-1.184	0.03	1.847
Item 4	2.599	-0.359	0.504	1.427	2.051	2.938	3.438	4.073	4.605
Item 5	4.057	-2.131	-0.833	0.294	1.299	2.068	2.651	3.595	4.605
Item 6	2.38	-3.9	-3.278	-2.588	-2.068	-1.336	-0.813	0.036	1.136
Item 7	2.951	-4.104	-2.662	-0.81	0.159	0.976	1.731	2.528	5.031
Item 8	2.891	-5.761	-4.299	-2.613	-1.425	-0.565	0.283	1.022	3.723
Item 9	3.416	-6.44	-4.873	-3.079	-1.884	-0.917	0.047	0.788	3.317

Note. Item discrimination is denoted as "a". b_1 to b_8 signifies item locations and reflects the threshold level of the latent trait necessary to have at least a 50% probability of endorsing the next response choice. Item parameters reflect those obtained with the GRM model.

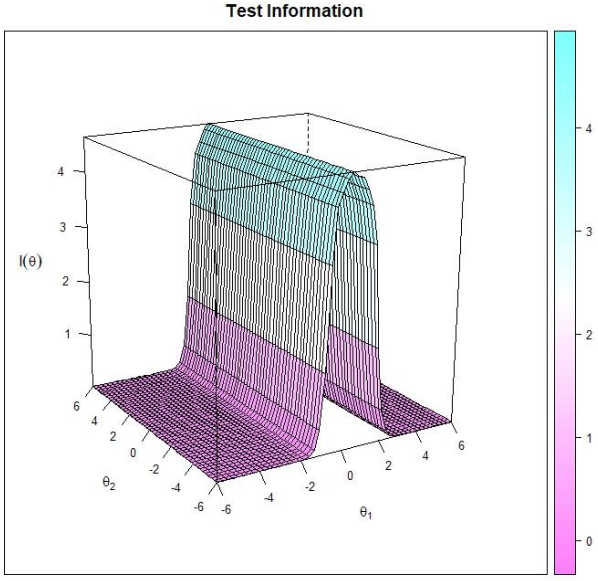
Information Plot for Item 1: Men



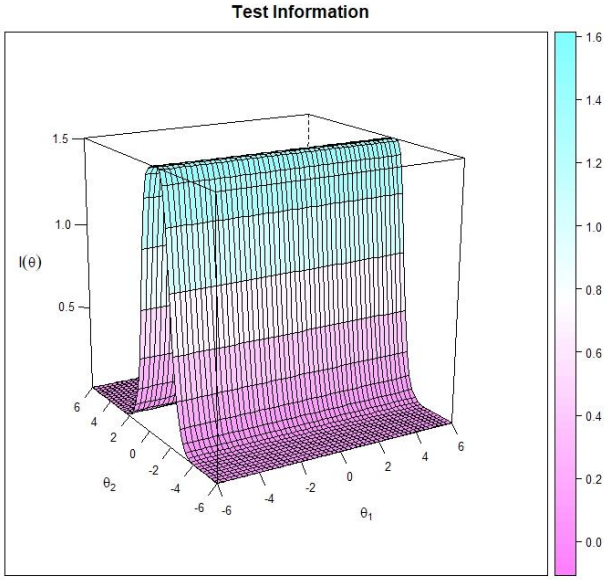
Information Plot for Item 2 : Men



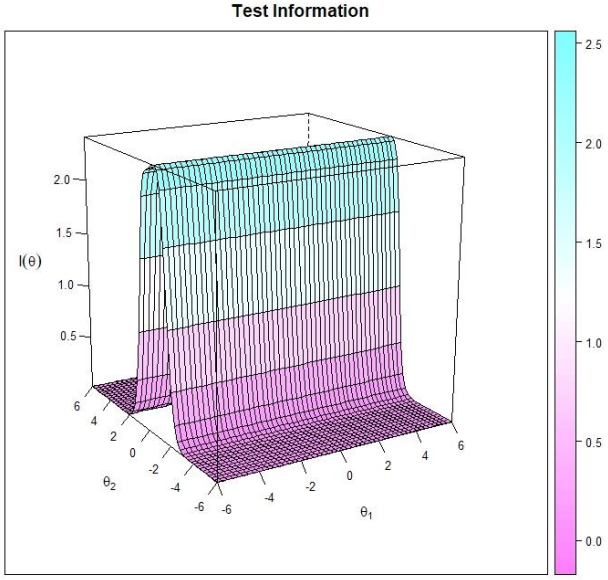
Information Plot for Item 3 : Men



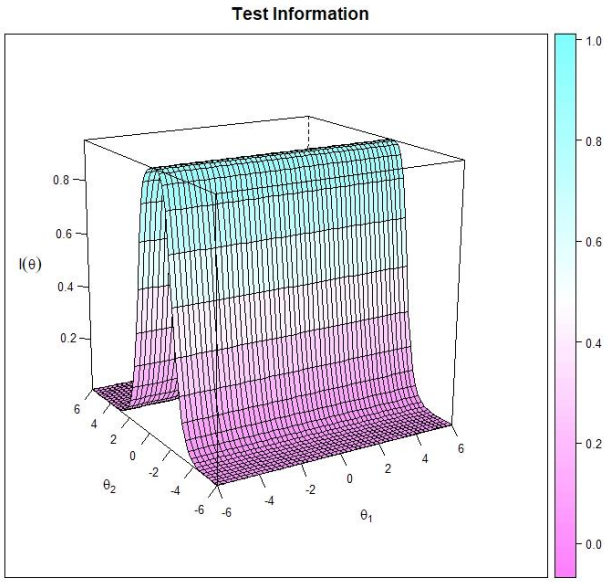
Information Plot for Item 4: Men



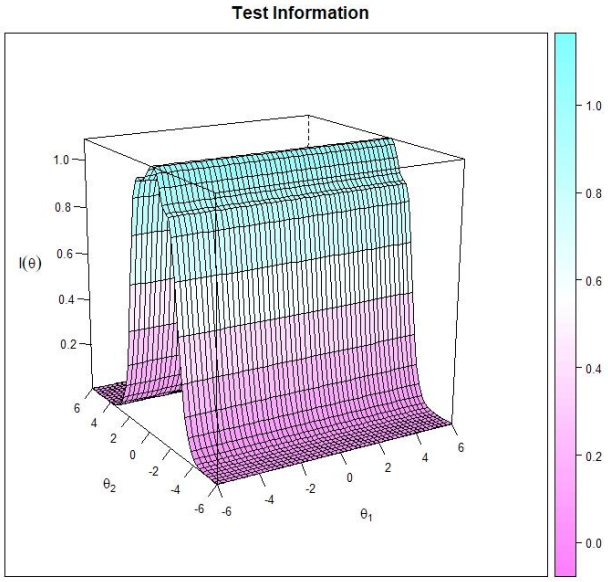
Information Plot for Item 5: Men



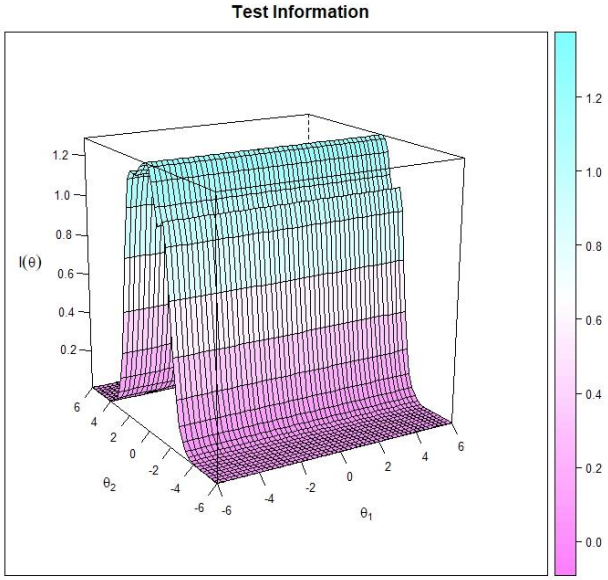
Information Plot for Item 6: Men



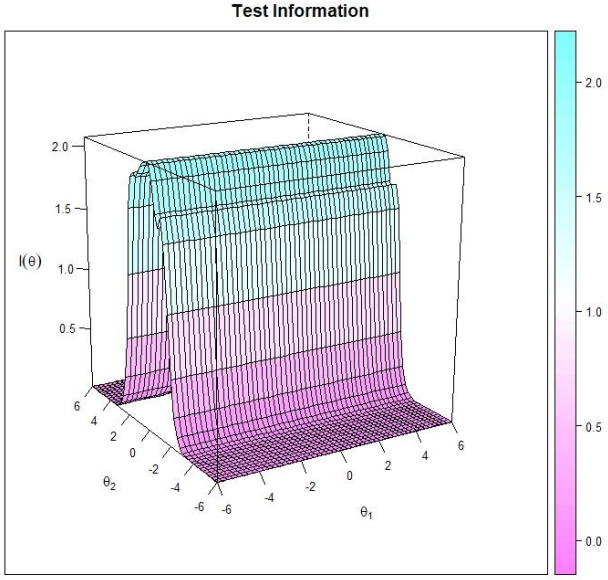
Information Plot for Item 7: Men



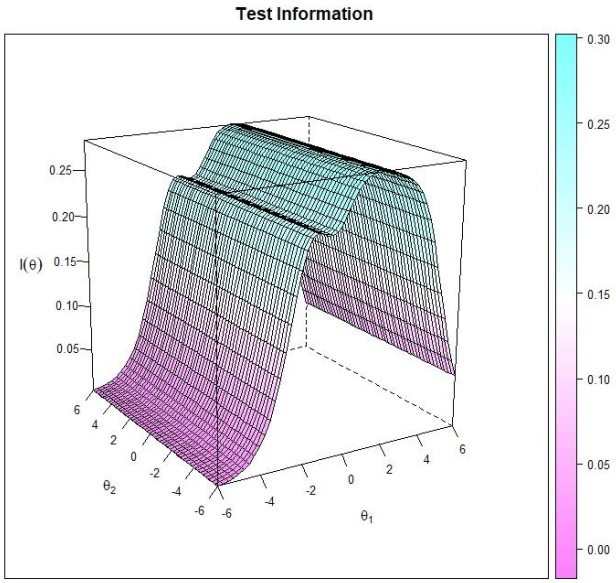
Information Plot for Item 8: Men



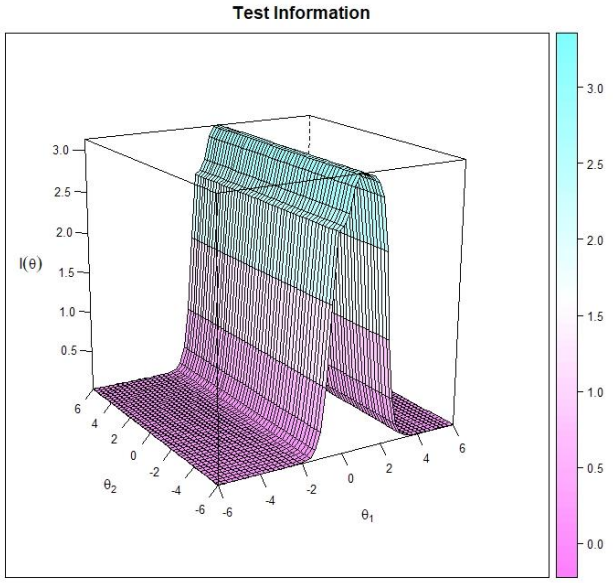
Information Plot for Item 9: Men



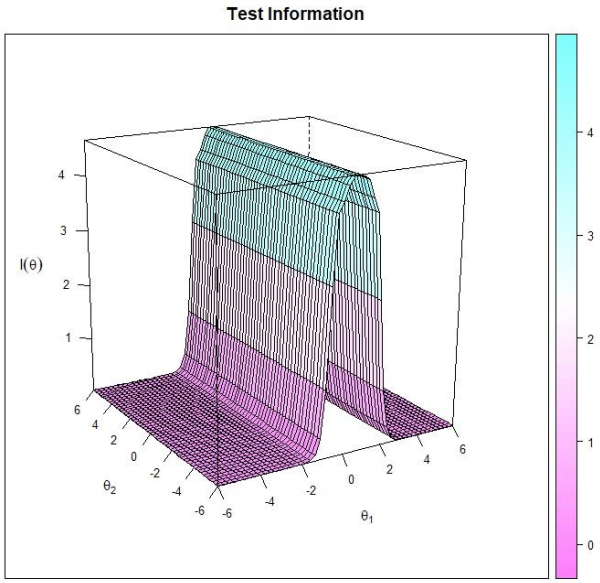
Information Plot for Item 1: Women



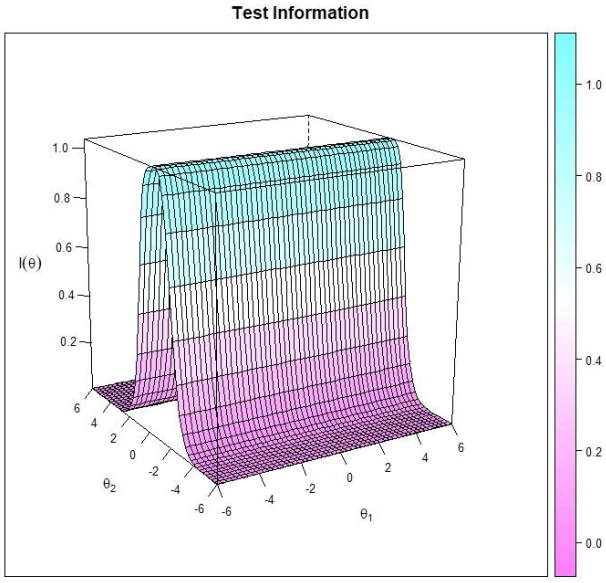
Information Plot for Item 2: Women



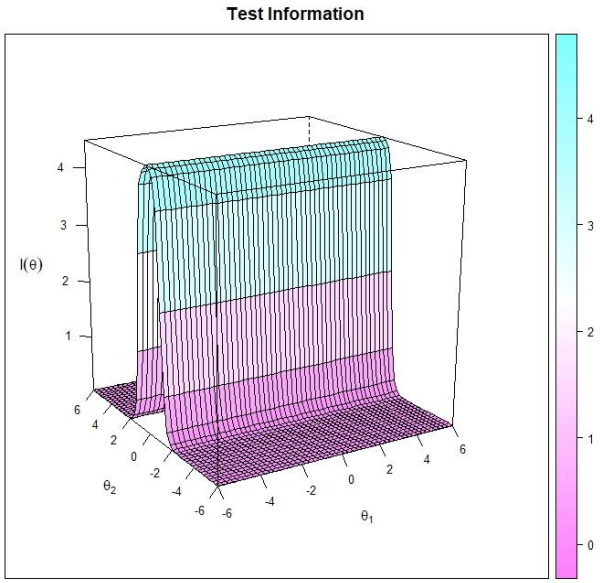
Information Plot for Item 3: Women



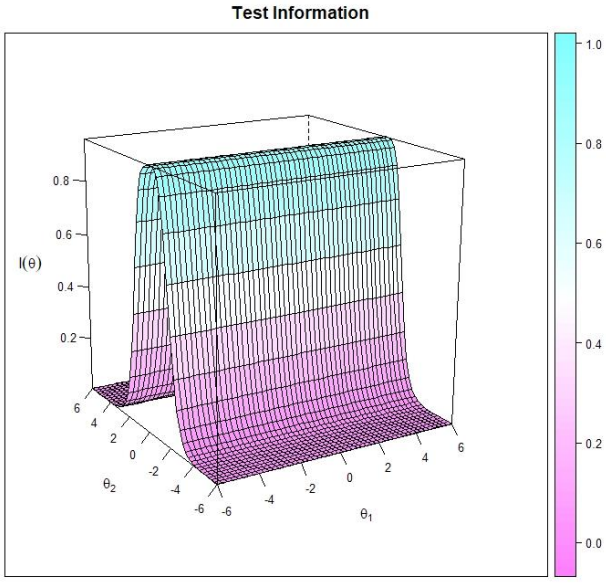
Information Plot for Item 4: Women



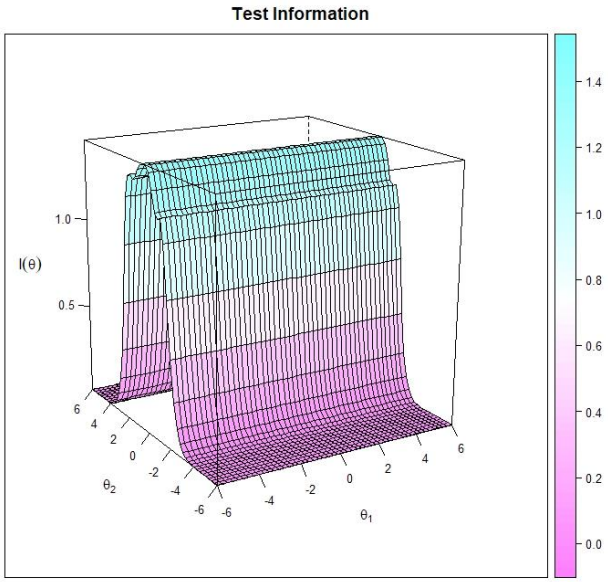
Information Plot for Item 5: Women



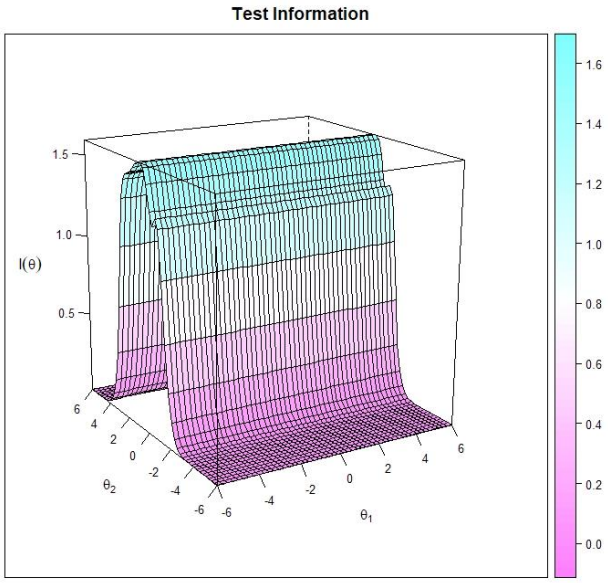
Information Plot for Item 6: Women



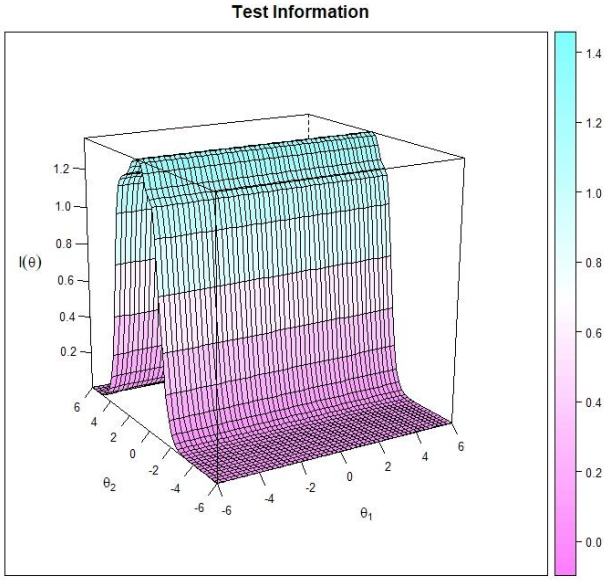
Information Plot for Item 7: Women



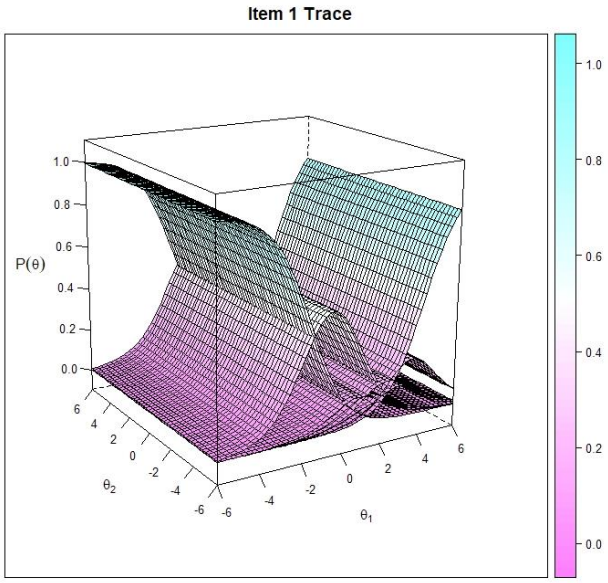
Information Plot for Item 8: Women



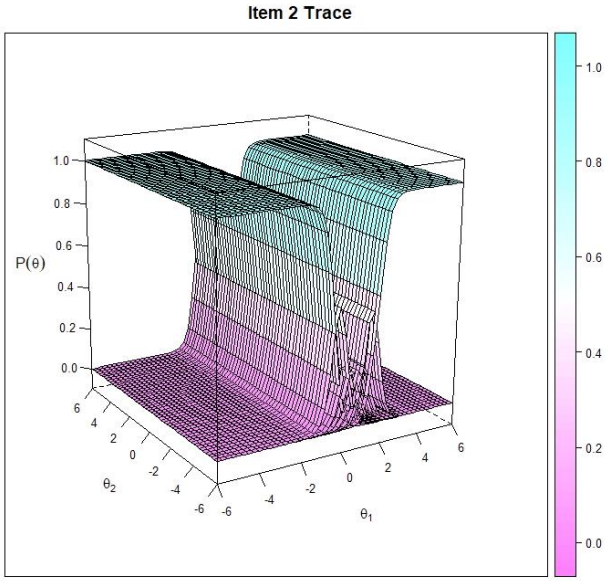
Information Plot for Item 9: Women



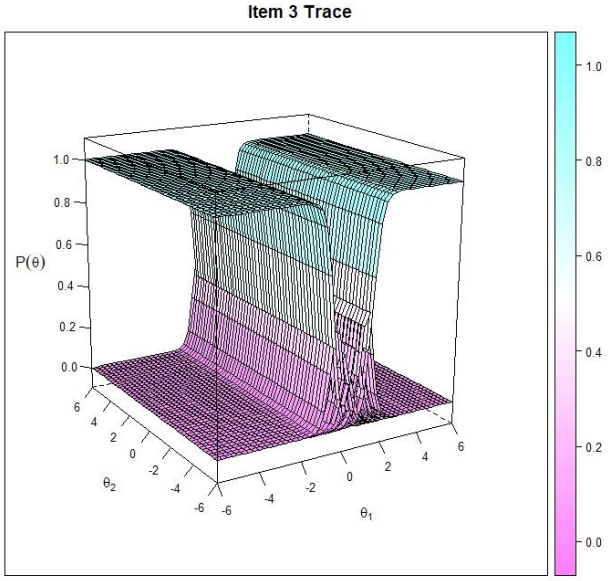
Item Characteristic Curves for Item 1: Men



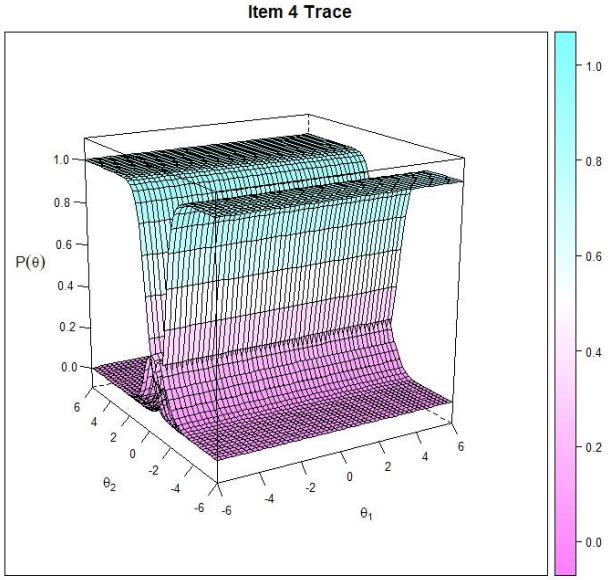
Item Characteristic Curves for Item 2: Men



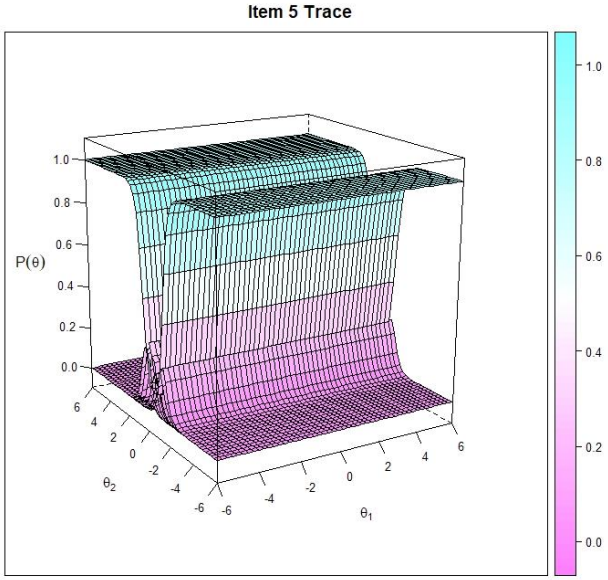
Item Characteristic Curves for Item 3: Men



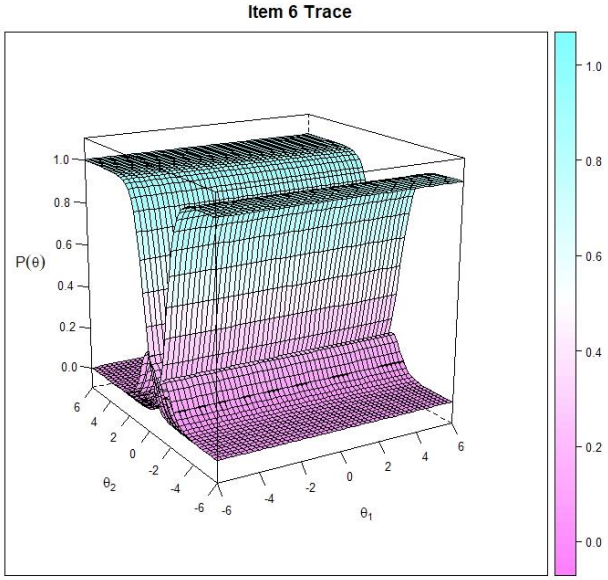
Item Characteristic Curves for Item 4: Men



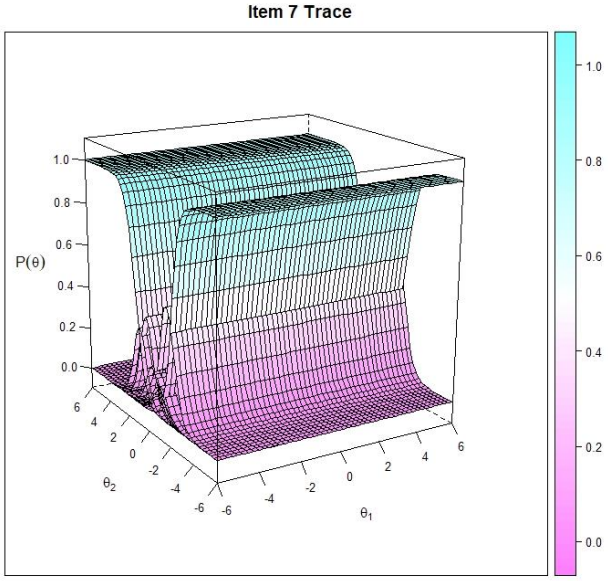
Item Characteristic Curves for Item 5: Men



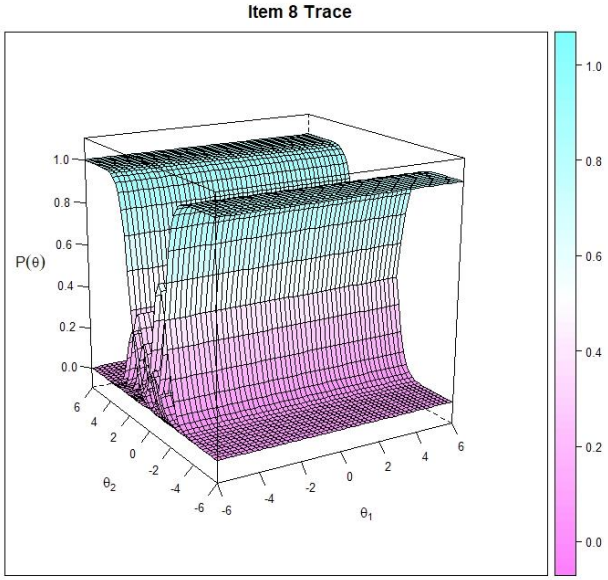
Item Characteristic Curves for Item 6: Men



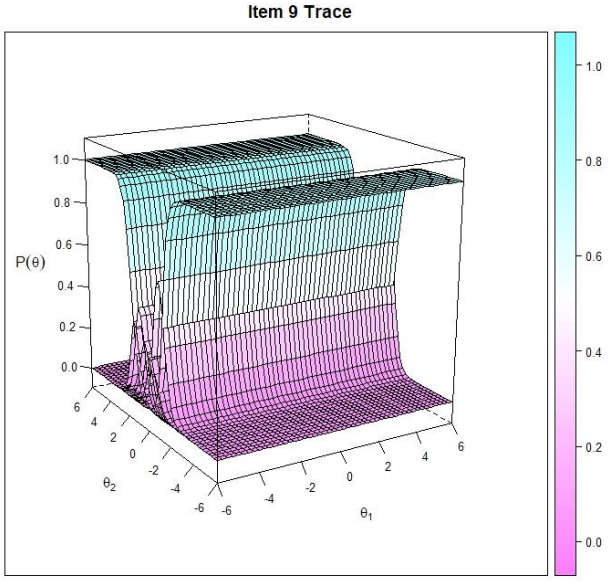
Item Characteristic Curves for Item 7: Men



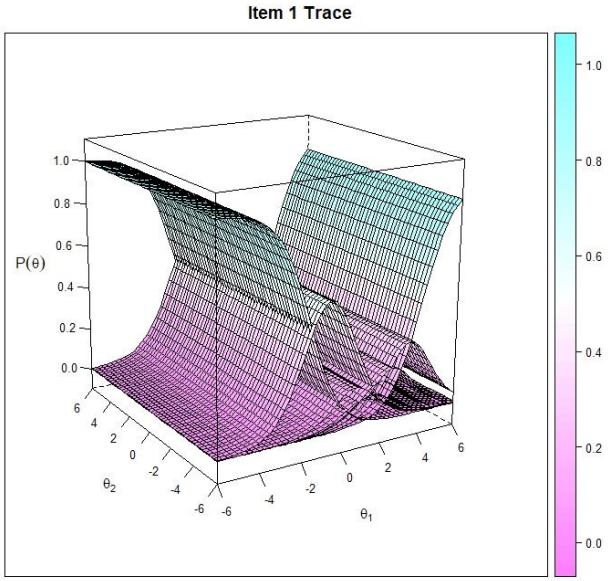
Item Characteristic Curves for Item 8: Men



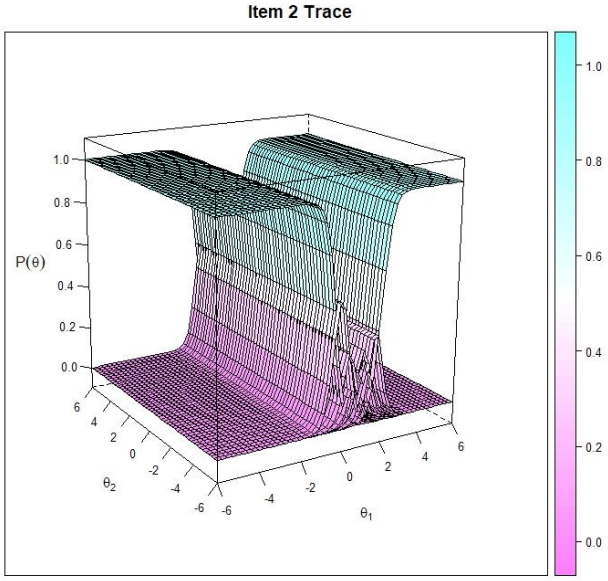
Item Characteristic Curves for Item 9: Men



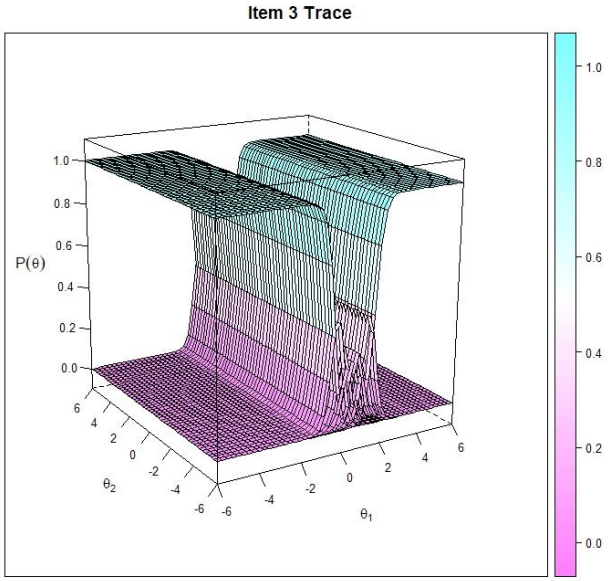
Item Characteristic Curves for Item 1: Women



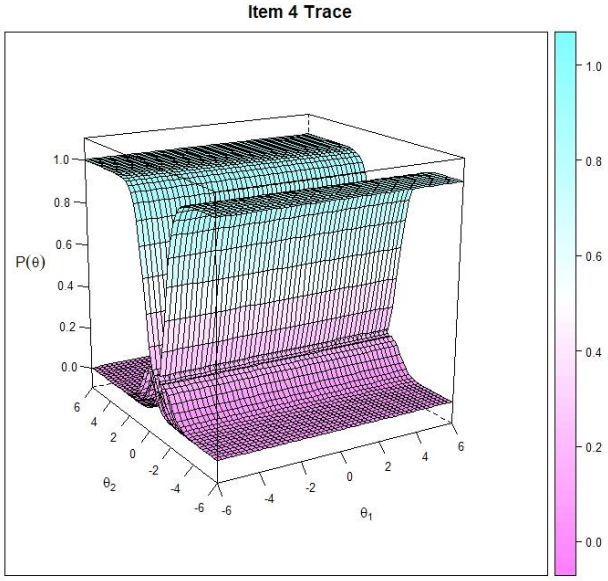
Item Characteristic Curves for Item 2: Women



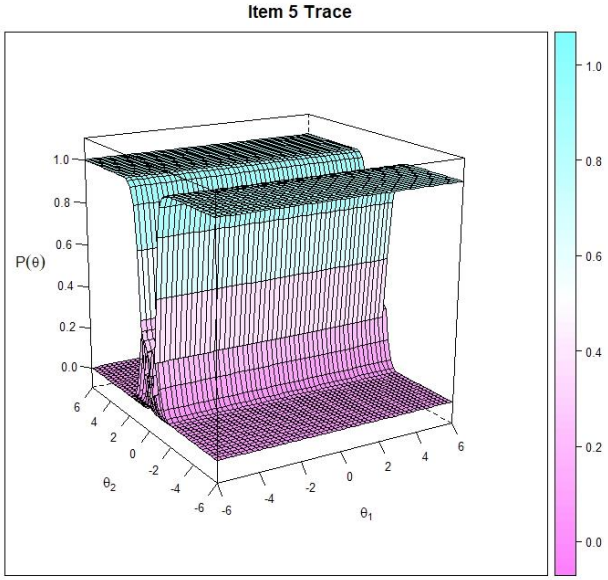
Item Characteristic Curves for Item 3: Women



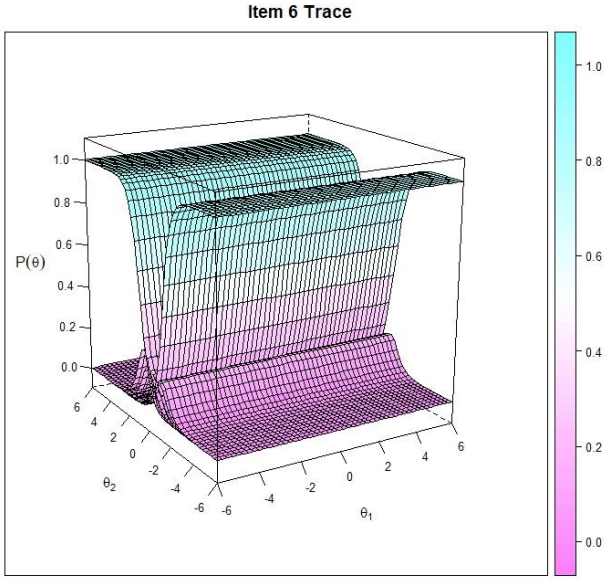
Item Characteristic Curves for Item 4: Women



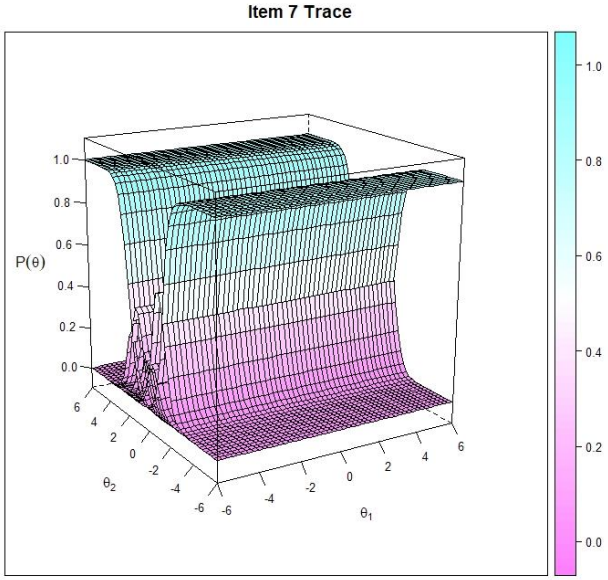
Item Characteristic Curves for Item 5: Women



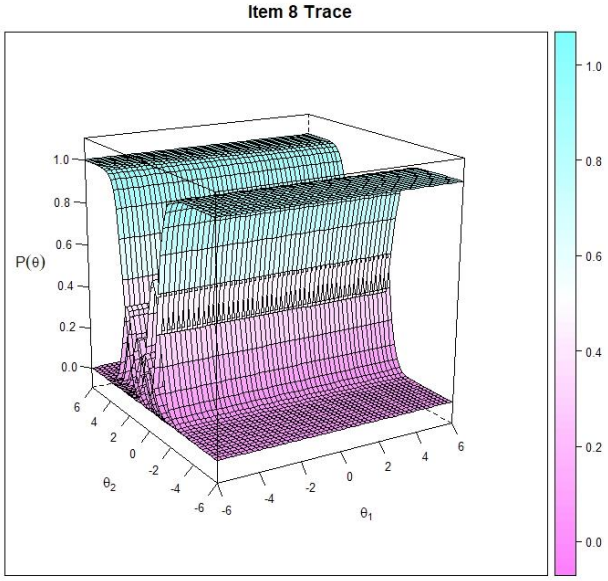
Item Characteristic Curves for Item 6: Women



Item Characteristic Curves for Item 7: Women



Item Characteristic Curves for Item 8: Women



Item Characteristic Curves for Item 9: Women